Genome Biology and Evolution

Comparative Genomic Analyses of the *Moraxella catarrhalis* Serosensitive and Seroresistant Lineages Demonstrate Their Independent Evolution

Joshua P. Earl[1,2,3,†], Stefan P.W. de Vries[4,5,†], Azad Ahmed[3], Evan Powell[3], Matthew P. Schultz[3], Peter W.M. Hermans[5], Darryl J. Hill[6], Zhemin Zhou[6], Crystala I. Constantinidou[7], Fen Z. Hu[1,2,3,8], Hester J. Bootsma[5,9,*], and Garth D. Ehrlich[1,2,3,8,*]

[1]Department of Microbiology and Immunology, Drexel University College of Medicine, Philadelphia, PA

[2]Center for Genomic Sciences and Center for Advanced Microbial Processing, Institute of Molecular Medicine and Infectious Disease, Drexel University College of Medicine, Philadelphia, PA

[3]Center for Genomic Sciences, Allegheny-Singer Research Institute, Allegheny General Hospital, Pittsburgh, PA

[4]Present address: Department of Veterinary Medicine, University of Cambridge, Cambridge, United Kingdom

[5]Laboratory of Pediatric Infectious Diseases, Radboud University Medical Centre, Nijmegen, The Netherlands

[6]Warwick Medical School, University of Warwick, Coventry, United Kingdom

[7]School of Cellular and Molecular Medicine, University of Bristol, Bristol, United Kingdom

[8]Department of Otolaryngology Head and Neck Surgery, Drexel University College of Medicine, Philadelphia, PA

[9]Present address: Centre for Infectious Diseases Research, Diagnostics and Screening, Centre for Infectious Diseases Control, National Institute of Public Health and the Environment (RIVM), Bilthoven, The Netherlands

[†]These authors contributed equally to this work.

*Corresponding author: E-mail: garth.ehrlich@drexelmed.edu; hester.bootsma@rivm.nl.

Accepted: February 18, 2016

Abstract

The bacterial species *Moraxella catarrhalis* has been hypothesized as being composed of two distinct lineages (referred to as the seroresistant [SR] and serosensitive [SS]) with separate evolutionary histories based on several molecular typing methods, whereas 16S ribotyping has suggested an additional split within the SS lineage. Previously, we characterized whole-genome sequences of 12 SR-lineage isolates, which revealed a relatively small supragenome when compared with other opportunistic nasopharyngeal pathogens, suggestive of a relatively short evolutionary history. Here, we performed whole-genome sequencing on 18 strains from both ribotypes of the SS lineage, an additional SR strain, as well as four previously identified highly divergent strains based on multilocus sequence typing analyses. All 35 strains were subjected to a battery of comparative genomic analyses which clearly show that there are three lineages—the SR, SS, and the divergent. The SR and SS lineages are closely related, but distinct from each other based on three different methods of comparison: Allelic differences observed among core genes; possession of lineage-specific sets of core and distributed genes; and by an alignment of concatenated core sequences irrespective of gene annotation. All these methods show that the SS lineage has much longer interstrain branches than the SR lineage indicating that this lineage has likely been evolving either longer or faster than the SR lineage. There is evidence of extensive horizontal gene transfer (HGT) within both of these lineages, and to a lesser degree between them. In particular, we identified very high rates of HGT between these two lineages for ß-lactamase genes. The four divergent strains are *sui generis*, being much more distantly related to both the SR and SS groups than these other two groups are to each other. Based on average nucleotide identities, gene content, GC content, and genome size, this group could be considered as a separate taxonomic group. The SR and SS lineages, although distinct, clearly form a single species based on multiple criteria including a large common core genome, average nucleotide identity values, GC content, and genome size. Although neither of these lineages arose from within the other based on phylogenetic analyses, the question of how and when these lineages split and then subsequently reunited in the human nasopharynx is explored.

Key words: comparative genome analysis, bacteria, Moraxella catarrhalis, evolution, phylogeny, distributed genome hypothesis, supragenome, pan genome.

Introduction

Moraxella catarrhalis is a human-restricted opportunistic respiratory tract pathogen within the pseudomonadales order of the gammaproteobacteria (www.metalife.com). This Gram-negative bacterium frequently colonizes the upper respiratory tract of young children, with two-thirds being colonized in their first year of life (Faden 2001). Although asymptomatic colonization is common, *M. catarrhalis* is an important etiological agent of upper respiratory tract infections in children; it is the third major cause of otitis media after *Streptococcus pneumoniae* and *Haemophilus influenzae* (Post et al. 1995; Aul et al. 1998; Dingman et al. 1998; Verduin et al. 2002; Murphy and Parameswaran 2009). Furthermore, it is the second most common cause of infectious exacerbations of chronic obstructive pulmonary disease, accounting for ~10% (2–4 million) of exacerbations annually in the United States (Faden 2001; Murphy et al. 2005). In rare instances it has been associated with bacteremia and septic arthritis (Melendez and Johnson 1991). The vast majority of *M. catarrhalis* strains are penicillin resistant (Hoban et al. 2001) and can form biofilms (Hall-Stoodley et al. 2006; Nistico et al. 2009) which together makes them highly recalcitrant to standard antibiotic treatment (Borriello et al. 2006; Perez et al. 2014).

Several molecular phylogenetic-typing methods, including restriction fragment length polymorphism analyses and multilocus sequence typing (MLST), have suggested that the *M. catarrhalis* species is composed of two main lineages (Bootsma et al. 2000a; Pingault et al. 2007; Wirth et al. 2007), but that divergent strains with limited homology also exist (Wirth et al. 2007). The first described phylogenetic lineage, often referred to as the seroresistant (SR) lineage, contains 16S ribotype (RB) 1 strains and displays a more pathogenic profile as the vast majority of its members are complement resistant and adhere efficiently to respiratory epithelial cells (de Vries et al. 2013, 2014). The other lineage, also known as the serosensitive (SS) lineage, consists of strains with RB2 and RB3, and mainly harbors *M. catarrhalis* isolates that are sensitive to complement-mediated killing and show less efficient adherence to respiratory epithelial cells (Bootsma et al. 2000a; Pingault et al. 2007; Wirth et al. 2007). Importantly, the two lineages differ significantly in their association with disease, with 51% of isolates from the SR lineage isolated from diseased individuals, while only 15% of the SS isolates were from diseased individuals (Verhaegh et al. 2008). Wirth et al. also demonstrated using MLST that the SS lineage is genetically more diverse than the SR lineage, and may represent remnants of an ancestral *M. catarrhalis* population. Their phylogenetic analysis indicated that the two lineages were genetically separated for an extended period. Additionally, they concluded that the genetic variation in the SS lineage appeared to be mainly the result of point mutation events, whereas SR lineage displayed increased mosaicism due to homologous recombination (Wirth et al. 2007).

Comparative whole-genome analyses of a collection of 12 clinically diverse SR isolates, each of different MLST, showed only limited genetic diversity (de Vries et al. 2010; Davie et al. 2011; Zomer et al. 2012) as compared with other nasopharyngeal pathogens including *H. influenzae* and *S. pneumoniae* (Shen et al. 2005, 2006; Hiller et al. 2007; Hogg et al. 2007; Donati et al. 2010; Boissy et al. 2011). In this study, we examined the evolution of the *M. catarrhalis* species through whole genome sequencing (WGS) of members of the SS lineage and a divergent lineage (Wirth et al. 2007) and compared these results with those obtained for the SR lineage strains (Davie et al. 2011).

Materials and Methods

Moraxella catarrhalis Strains and Growth Conditions

The *M. catarrhalis* SS lineage isolates and the highly divergent strains sequenced for this study (Bootsma et al. 2000a; Wirth et al. 2007; Hill et al. 2012; Stol et al. 2012) and the previously sequenced SR lineage isolates (de Vries et al. 2010; Davie et al. 2011; Zomer et al. 2012) are listed in table 1. *Moraxella catarrhalis* strains were routinely cultured on brain heart infusion (BHI) agar plates at 37 °C in a humidified atmosphere containing 5% CO2, or in BHI broth at 37 °C in an atmosphere containing 5% CO_2 with agitation (200–250 rpm). All strains including the divergent strains (Vaneechoutte et al. 1990) were originally identified at the time of their isolation or characterization as Gram-negative diplococci, DNAse positive, oxidase positive, 4-methylumbelliferyl butyrate (Vaneechoutte et al. 1988) positive, and glucose negative. For DNA isolations, strains were cultured until early exponential phase growth (OD_{620nm} [optical density] of 0.2–0.4); chromosomal DNA was isolated using Genomic-tip 20/G columns (Qiagen, Venlo, Netherlands) according to manufacturer's instructions or as described (Marshall et al. 1997).

Genome Sequencing Assembly and Annotation

WGS was performed for 15 SS strains as described (Hiller et al. 2007; Hogg et al. 2007; Donati et al. 2010; Boissy et al. 2011; Davie et al. 2011; Ahmed et al. 2012) using a 454 LifeSciences FLX+ genome sequencer (Roche, Branford, CT) with Titanium chemistry. Two additional SS strains were sequenced using Illumina HiSeq and one SS strain was sequenced on a Pacific Biosciences RSII system using SMRT C1 chemistry according to the manufacturer's instructions (Menlo Park, CA). The four

Table 1

Moraxella catarrhalis Serosensitive and Seroresistant Lineage Isolates: Strain and Molecular-Typing Data

Strain	Source	Location	Reference(s)	Accession No.	Lineage	Ribo Type	MLST Type	*abcZ*	*adk*	*efp*	*fumC*	*glyRS*	*mutY*	*ppa*	*trpE*	β-Lactamase	LOS Type
A9	NA	Africa	Bootsma et al. (2000)	SAMN04122787	SeroS	2	New	38	41	9	new	64	new	new	1	None	C
C031	Nose, OM, child	Netherlands	Stol et al. (2012)	SAMN04122788	SeroS	2	ST1	1	1	1	1	1	1	1	1	*bro-1*	A
C10	Sputum, COPD patient, adult	Netherlands	This study	SAMN04122789	SeroS	2	NP-ST	11	10	8	12	11	new	11	1	None	A
F18	Sputum, LRT infection	Netherlands	Bootsma et al. (2000)	SAMN04122790	SeroS	3	ST170	45	37	15	5	60	49	4	12	*bro-1*	A
F20	Sputum, LRT infection	Netherlands	Bootsma et al. (2000)	SAMN04122791	SeroS	2	ST219	10	10	1	12	11	11	10	6	None	A
F21	Sputum, LRT infection	Netherlands	Bootsma et al. (2000)	SAMN04122792	SeroS	2	ST45	20	21	14	19	12	23	19	10	*bro-1*	A
F23	Sputum, LRT infection	Netherlands	Bootsma et al. (2000)	SAMN04122793	SeroS	2	ST46	21	4	7	12	25	24	20	11	*bro-1*	A
F24	Sputum, LRT infection	Netherlands	Bootsma et al. (2000)	SAMN04122795	SeroS	3	ST169	38	4	27	5	59	25	40	1	*bro-1*	A
N1	Nose, child	Netherlands	Bootsma et al. (2000)	SAMN04122796	SeroS	2	ST37	19	15	13	17	23	20	18	3	None	A
N12	Nose, child	Netherlands	Bootsma et al. (2000)	SAMN04122797	SeroS	2	ST22	15	4	11	5	16	16	15	3	None	A
N15	Nose, child	Netherlands	Bootsma et al. (2000)	SAMN04122798	SeroS	3	ST23	16	15	1	15	12	17	16	9	*bro-1*	A
N19	Nose, child	Netherlands	Bootsma et al. (2000)	SAMN04122799	SeroS	2*	New	30	1	1	1	1	new	1	1	None	A
N4	Nose, child	Netherlands	Bootsma et al. (2000)	SAMN04122800	SeroS	3	ST23	16	15	1	15	12	17	16	9	*bro-1*	A
N6	Nose, child	Netherlands	Bootsma et al. (2000)	SAMN04122801	SeroS	2	ST1	1	1	1	1	1	1	1	1	None	A
R4	NA	Netherlands	Bootsma et al. (2000)	SAMN04122802	SeroS	3	ST48	23	4	15	5	26	25	4	12	*bro-1*	A
Z18	Pharynx, laryngitis, adult	Sweden	Bootsma et al. (2000)	SAMN04122845	SeroS	2	ST52	15	4	11	5	28	16	22	1	None	A
Mx1	Sputum	United Kingdom	This study	SAMN04122860	SeroS	3	New	45	new	15	5	60	49	4	12	*bro-1*	A
S11	Nasopharynx	Sweden	This study	SAMN04122862	SeroS	2	New	23	41	11	5	new	24	new	new	*bro-2*	A
BBH18	Sputum, COPD exacerbation	Denmark	de Vries et al. (2010)	CP002005	SeroR	1	128	8	30	2	7	32	3	3	2	*bro-2*	B
7169	Middle ear, OME, child	United States	Davie et al. (2011)	AERC00000000	SeroR	1	82	3	18	3	4	3	9	3	2	*bro-1*	B
103P14B1	COPD exacerbation, adult	United States	Davie et al. (2011)	AERE00000000	SeroR	1	187	8	26	2	3	2	22	41	2	*bro-1*	A
12P80B1	COPD exacerbation, adult	United States	Davie et al. (2011)	AERG00000000	SeroR	1	185	50	25	12	3	37	52	3	2	*bro-2*	C
46P47B1	COPD exacerbation, adult	United States	Davie et al. (2011)	AERF00000000	SeroR	1	186	3	20	2	7	62	15	8	2	None	B
O35E	OME, child	United States	Davie et al. (2011)	AERL00000000	SeroR	1	146	2	6	2	4	57	22	8	2	*bro-1*	A
BC1	Tracheal aspirate, bronchiolitis, child	United States	Davie et al. (2011)	AERH00000000	SeroR	1	216	25	18	3	4	6	9	3	2	*bro-1*	A
BC7	Middle ear, OME, child	United States	Davie et al. (2011)	AERI00000000	SeroR	1	217	8	18	12	3	29	3	3	2	*bro-1*	A
BC8	Sinus wash, sinnusitis, child	United States	Davie et al. (2011)	AERJ00000000	SeroR	1	162	2	2	12	7	20	6	3	2	*bro-1*	C
C072	Middle ear fluid, OME, child	Netherlands	Davie et al. (2011)	AERK00000000	SeroR	1	199	8	8	2	3	2	3	9	2	*bro-1*	A
101P30B1	COPD exacerbation, adult	United States	Davie et al. (2011)	AEPC00000000	SeroR	1	218	8	26	2	3	2	3	25	2	*bro-1*	A
RH4	Blood	Denmark	Zomer et al. (2012)	AMSO00000000	SeroR	ND	New	new	6	3	2	20	9	3	2	*bro-1*	A
43617	Transtracheal aspirate, chronic bronchitis, coal miner	Belgium	Wang et al. (2007)	AX067426	SeroR	1	25	3	3	3	4	18	3	3	2	*bro-2*	B
Z7574	NA	United States	Wirth et al. (2007)	SAMN04122865	Div	NA	127	37	29	21	26	51	42	32	15	NA	NT
Z7547	NA	Ethiopia	Wirth et al. (2007)	SAMN04122866	Div	NA	116	33	29	19	22	45	38	30	14	NA	NT
Z7542	NA	Ethiopia	Wirth et al. (2007)	SAMN04122867	Div	NA	116	33	29	19	22	45	38	30	14	NA	NT
Z7546	NA	Ethiopia	Wirth et al. (2007)	SAMN04122868	Div	NA	116	33	29	19	22	45	38	30	14	NA	NT

NOTE.—OM = otitis media; COPD = chronic obstructive pulmonary disease; LRT = lower respiratory tract; OME = chronic otitis media with effusion; NA = not available; SeroS = sero-sensitive; SeroR = Sero-resistant; Div = divergent; ND = not determined; New* = a novel MLST based on novel alleles will need to be assigned; new* = a novel ST allele based on WGS of the strain will need to be assigned; *currently the *Moraxella catarrhalis* MLST database does not accept sequence data from non-Illumina sequenced strains, thus it was not possible to get a novel ST assigned in these cases; NT = nontypeable.

divergent strains identified by Wirth et al. (2007) were sequenced with an Illumina MiSeq (Illumina, San Diego, CA). Following sequencing, the number of reads, the average read lengths, the depth of coverage, the GC content, and the number of contigs for each strain was determined. Sequencing with the 454 and Pacific Biosciences RS systems was conducted at the Center for Genomic Sciences (CGS) at the Allegheny-Singer Research Institute. As previously described (Davie et al. 2011; Ahmed et al. 2012), the 454 raw sequence reads for each strain were assembled into contigs using the Roche/454 Life Sciences' GS de novo Newbler assembler (version 2.0.00.20) using the default parameters except for minimum overlap identity which was adjusted to obtain the fewest contigs. For all 454-based assemblies, we called single nucleotide polymorphisms (SNPs) against the self-assembly. The C10 strain, sequenced on the RS II system, was assembled using the Celera WGS-assembler (Myers et al. 2000). Illumina-sequenced strains were assembled with either Newbler (v2.6) or Ray (v2.3.1). The final genome assemblies have been deposited in GenBank, and the corresponding accession numbers are provided in table 1. The assembled genomes were submitted to RAST (Rapid Annotation using Subsystems Technology; http:///RAST.nmpdr.org) for automated annotation (Aziz et al. 2008). Specific gene functions were inferred to be present in a genome if any region of that genome was inferred to be homologous to a region in another genome (see below) with the annotated function (Hogg et al. 2007).

Gene Clustering and Comparative Genome Analyses

The CGS comparative genomics pipeline (Hiller et al. 2007; Hogg et al. 2007; Donati et al. 2010; Boissy et al. 2011; Davie et al. 2011; Ahmed et al. 2012) was used to take the RAST-derived annotated coding sequences and perform gene cluster analyses. Briefly, gene annotation sequences from every strain were collated into two multi-fasta files, both as nucleic acid (NA), and amino acid (AA) sequences. All-against-all alignments between NA and AA were done with the program tfasty in the Fasta v3.6 package (Pearson and Lipman 1988). In addition, to capture instances where the nucleotide sequence of a gene exists in a genome, but was not annotated by RAST, the nucleotide sequences are also aligned against the genome contigs with the fasta program. A gene was defined as being a member of a cluster if it had a Fasta alignment with a percent identity of at least 70% the length of the shorter sequence to at least one other gene/sequence (annotated/unannotated) in the cluster (single-linkage) (Hogg et al. 2007). The gene clusters were then binned as either core or distributed based on their possession profiles. Interstrain comparisons were then performed for all 595 ([34 × 35]/2) possible strain pairs with regard to the number of gene possession similarities and differences. This results in a set of gene clusters, each of which is either present or absent in each genome, and may have multiple representatives from a given genome; clusters represented in all genomes are designated "core" and those remaining are designated "distributed."

Whole-genome alignments of all 35 *M. catarrhalis* genomes (13 SR, 18 SS, and 4 divergent *M. catarrhalis* isolates), as well as an *Moraxella macacae* genome used as a phylogenetic outgroup were performed using default parameters in Mauve v2.3.1 (Darling et al. 2004). We also performed analyses of the SR and SS lineages using the divergent *M. catarrhalis* strains as an outgroup. For each genome, its locally collinear blocks (LCBs), as identified by Mauve, were concatenated together and referred to as the "concatenated core" sequence, which is independent of gene annotations. These sequences were aligned with Mafft v7.127b with default "fast" parameters. A network was inferred by importing the alignment directly into SplitsTree4 (Huson and Bryant 2006), which automatically calculates the network. Additionally, this alignment was also used as input to the program RAxML v8.0.20 (Stamatakis 2006) wherein the GTGRAMMA model was used and one thousand bootstraps were performed to measure tree branching consistency as follows: raxlHPC-PTHREADS-SSE3 –f a –s input.fasta –n boot –m GTRGAMMA –d 1234 –x 1234 -#1000 –T 16 –n T20 –o output

Genomic Recombination Analyses

Moraxella catarrhalis is a naturally competent species. Although it is possible to explain reticulation within the phylogenetic networks as indicative of recombination events (Huson and Bryant 2006), we also explored an additional analysis to provide more rigorous evidence of horizontal gene transfer (HGT). To investigate the number of regions within these genomes potentially undergoing HGT, the recently published orderedPainting method (Yahara et al. 2014) was used. This method infers "hotspot" regions from a multiple genome alignment by examining SNPs. All SNPs containing no gap or ambiguous characters were extracted from the concatenated core genome alignment (as described above) by a custom Ruby script and analyzed on our cluster using the orderedPainting software. From these data, we identified the recombination hotspots from which we built phyml trees. First, we used JmodelTest V 2.1.6 to calculate likelihood using all 88 possible different DNA model combinations. Using the AIC criteria, the best scoring model was found to be GTR + I + G. Phyml version 20120412 was then used to calculate the maximum-likelihood tree using the command line generated by the jModelTest program: pyml –i/align.phy –d nt –n 1 –b 0 –run_id GTR + I+G –m 012345 –f m –v e –c 4 –a e –no_memory_check –o tlr –s NNI

Strain Groupings

The allelic and gene possession neighbor grouping networks were calculated from distance matrices as described (Hall et al.

2010). Briefly, allelic pairwise distance was calculated as one minus the average percent identity of all aligned core genes, and the gene possession pairwise distance was calculated as one minus the total number of distributed genes either both present or both absent in each genome, divided by the total number of distributed genes in the supragenome. Networks were generated for these analyses in SplitsTree4 by importing the distance matrices which automatically computed the network image (Huson and Bryant 2006).

Supragenome Modeling

Modeling of the lineage-specific and whole species supragenomes was performed as described using the modified Finite Supragenome Model (FSGM; Hogg et al. 2007; Boissy et al. 2011).

Age Inference Modeling

We used BEAST v1.75, a Bayesian-based model utilizing Markov chain Monte Carlo methods to calculate the age of evolutionary divergence between the SR and SS *M. catarrhalis* lineages (Drummond and Rambaut 2007). We used the concatenated core sequence alignment of *M. macacae* and *M. catarrhalis* as input and the generally accepted date of 23.5 Ma for the cercopithoid-hominoid split as the Bayesian prior for the date of the split at the tree root (Pickford and Andrews 1981) as we reasoned that the independent evolution of these two species would have tracked with the divergence of their host ancestors. A strict fixed clock model was required for age estimates, and the general time reversible (GTR) model was used, as from the limited models available in BEAST, it had the lowest AIC/BIC (Akaike information criterion/Bayesian information criterion) score as calculated from the alignment in the program Topali v2.5.

Phenotype–Genotype Association Mapping

The web-based tool PhenoLink (Bayjanov et al. 2012) was used to link observed serum resistance/sensitivity and adherence phenotypes to genome content. To this end, strains were divided into two classes for each examined phenotype as follows: Serum resistant (>10% survival) versus serum sensitive (<10% survival); intermediate to high adherent (>20%) versus low adherent (<20%). Subsequent association analyses were performed using default parameter settings.

In Silico Phylogenetic Analysis

The 16S ribosomal RNA (rRNA) type for the SS and SR strains was determined according to Bootsma et al. (2000a). Allelic sequences of the eight MLST genes were analyzed via http://mlst.warwick.ac.uk/mlst/dbs/Mcatarrhalis as detailed (Wirth et al. 2007). Lipooligosaccharide (LOS) serotyping was done according to Edwards et al. (2005), and typing of the beta-lactamase *bro* gene was performed according to Bootsma et al. (1996). An attempt was made to deposit the MLST data, but the curator informed us the site is down for reconfiguration.

Serum Resistance Assay

Moraxella catarrhalis isolates were cultured to mid-log phase (OD_{620nm} ~ 1.0). Cultures were washed in four volumes of phosphate buffered saline (PBS) supplemented with 0.15% gelatin (PBS-G) and resuspended in five volumes PBS-G supplemented with 1 mM $CaCl_2$ and 0.2 mM $MgCl_2$ (PBS-G Mg^{2+} Ca^{2+}). Resuspended bacterial cultures were mixed with an equal volume of 80% pooled normal human serum (NHS) (GTI diagnostics) diluted in PBS-G Mg^{2+} Ca^{2+} or NHS that was heat inactivated for 30 min at 56 °C (NHS-HI). Colony-forming units (CFUs) were determined at 0 and 60 min postincubation by plating 10-fold serial dilutions on BHI plates. The survival percentage was calculated relative to NHS-HI ($n \geq 3$) and statistical significance was determined using a Mann–Whitney test in GraphPad Prism 5.0.

Adhesion to Respiratory Tract Epithelial Cells

Adhesion of *M. catarrhalis* strains to the human pharyngeal epithelial cell line Detroit 562 (ATCC CCL-138) and the type II alveolar epithelial cell line A549 (ATCC CCL-185) was performed as detailed in de Vries et al. (2009). Briefly, cells were seeded in 24-well plates at 2×10^5 Detroit 562 cells/well 2 days prior to the experiment (at 24 h the medium was refreshed), or 4×10^5 A549 cells/well 1 day prior to the experiment. For both cell lines, monolayers of approximately 1×10^6 cells/well were used for adherence assays. *Moraxella catarrhalis* isolates were cultured to mid-log phase (OD_{620nm} ~ 1.0) and stored in the presence of 20% glycerol at −80 °C. Cells were washed twice with PBS before the bacteria were allowed to adhere to the epithelial cells (multiplicity of infection, 10:1) for a 1 h period in infection medium (DMEM-GlutaMAX™-I with 1% fetal calf serum). Thereafter, cells were washed three times with PBS to remove nonadherent bacteria. After detachment and lysis of the epithelial cells through addition of 1% saponin (Sigma-Aldrich, St. Louis, MO) in PBS-G, CFUs were enumerated by plating 10-fold serial dilutions on BHI plates. The percentage adherence ($n \geq 4$) was calculated as the percentage of the inoculum that bound to epithelial cells. Statistical significance was determined using a Mann–Whitney test in GraphPad Prism 5.0.

Results

Whole-Genome Sequencing of *Moraxella catarrhalis* Strains

Previous gene-based phylogenetic studies have suggested that the *M. catarrhalis* species contains at least two distinct groups, with the SR lineage strains forming a unique clade separate from the larger SS grouping(s). Phenotypically, the SR and SS

lineages have been shown to differ in their capacity to withstand the action of the human complement system and to adhere to host epithelial cell lines (Bootsma et al. 2000a; Pingault et al. 2007; Wirth et al. 2007; Verhaegh et al. 2008). In this study, using WGS and comparative genome analytics, we endeavored to determine the total clade structure of *M. catarrhalis* and any ancestral relationships among the various subpopulations identified. Previously, work by Davie et al. (2011) described the sequencing and comparative genomic analyses of 12 SR lineage strains, all of 16S ribotype 1 (RB1), which were selected on the basis of their divergent clinical and geographic sites of isolation. To gain further insight into the overall genetic diversity of the *M. catarrhalis* species as a whole, in this study 18 SS lineage strains, one additional SR lineage strain, and four divergent strains (which by MLST typing did not cluster with either the SR or SS lineages; Wirth et al. 2007) were subjected to whole-genome sequence analysis (table 1). The SS isolates were selected based on their genetic diversity, as determined both by 16S ribotypes {RB2 (n = 12) and RB3 (n = 6)} and MLST data (Bootsma et al. 2000; Pingault et al. 2007; Wirth et al. 2007), as well as for their clinical sites of recovery, which represented diverse niches from both diseased and healthy hosts. The sequencing platforms, assembly softwares, levels of coverage, numbers of contigs, genome size, and GC content for all 35 *M. catarrhalis* strains are presented in summary form (table 2).

The strain-specific genomic data were then analyzed comparatively to obtain average measures of diversity within and among the three groups of strains. This was accomplished by computing the average and range of both the percent GC base content and genome size (table 2), as well as determining the average nucleotide identity (ANI) (Konstantinidis et al. 2006) and tetranucleotide identity (Richter and Rosella-Mora 2009) using both BLAST and MUMmer algorithms. Between the SS and SR clades, the average ANI among all possible strain pairs was 95.78% with a range of 95.6–95.91% (supplementary tables S1 and S2, Supplementary Material online), and the average tetranucleotide identity was 99.92%, with a range 99.76–99.96%. In contrast, comparing all of the SS and SR strains collectively against the four divergent strains shows an average ANI of only 89.04% with a range of 87.98–89.68%, with an average tetranucleotide identity of 95.57% and a range of 95.29–95.83%. Similarly, comparing the WGS data among all strains showed that SR and SS lineages collectively span a percent GC spectrum of only 41.4–41.7%, while the four divergent strains have their own highly distinct GC content ranging very narrowly from 43.6% to 43.69% (table 2). With regard to genome sizes, the SR lineage displays a significantly smaller average genome size than the SS lineage, 1.89 versus 1.93 Mb (P = 0.0275, Mann–Whitney test), with both the RB2 and RB3 ribotypes within the SS lineage displaying essentially identical genome size averages. The range in genome sizes within each of the lineages is ~100 kb. In contrast, the divergent strains have a mean genome size of 2.16 Mb (again with a range of ~100 kb), making them 13% larger on average than the combined SS and SR strain set, and even larger than the *M. macacae* genome of 2.08 Mb. *Moraxella macacae* (Ladner et al. 2013), which we used as an outgroup for the phylogenetic tree, differed from the SS, SR, and divergent groups with an average ANI of 69.83% with a range of 69.00–70.31% while its genome had a third unique GC content of 39%. A MAUVE alignment of all 31 SS and SR strains and the *M. macacae* genome revealed that the overall genome structure across the genus is reasonably conserved (supplementary fig. S1, Supplementary Material online). There is, however, as discussed below, evidence of frequent HGT and moderate scale inversions, insertions, and deletions among all SS and SR strains.

We then calculated empirical values for the *M. catarrhalis* lineage-specific core genomes, distributed genomes, and supra(pan)genomes, as well as the species-wide values (both including and excluding the divergent strains; table 3 and fig. 1). Adding the four divergent strains to the combined core of the SS and SR lineages significantly reduced the size of the core genome from 2,318 to 2,041 genes, a >10% decrease. The observed supragenomes for the individual SR and SS lineages contained 3,240 and 3,524 genes, respectively, whereas the combined supragenome for these two lineages contained 3,854 genes. A massive increase in supragenome size to 5,684 genes was observed when the 4 divergent strains were included in the analysis. The number of distributed genes increased by 2,107 (from 1,536 to 3,643) with the addition of just four divergent strains to the combined SR/SS lineage supragenome. Interestingly, the vast majority of this increase (1,905 genes) occurred with the addition of the first divergent strain which resulted in more than doubling the total number of distributed genes in the supragenome with the addition of the next three divergent genomes only adding an additional 202 distributed genes. Collectively, these data are supportive of the divergent strains being *sui generis*.

These gene content data were then used as inputs into the revised FSGM (Hogg et al. 2007; Boissy et al. 2011) to model the total diversity present in each of the SR and SS lineages, the two lineages combined, and the species as a whole including the divergent strains. To compare similar numbers of strains for each of these groupings, we made predictions based on 50 strains (supplementary table S3, Supplementary Material online). For the SR and SS lineages, the FSGM predicts that the lineage-specific supragenomes are greater than 90% complete following the addition of 13 and 11 genomes, respectively, and the species-level (excluding the divergent strains) supragenome is 90% complete following the addition of 12 genomes. The FSGM, based on the actual number of genomes sequenced (N's = 13, 18, and 31, for the SR, SS, and species as a whole, respectively), gave supragenome size predictions of 3,240, 3,524, and 3,854 genes which are identical to the observed values. The only substantive difference

Table 2

Genome Characteristics of 35 *Moraxella catarrhalis* Strains

Strain	Bnum	Size (MB)	Coverage	Contigs	N50	% GC	Sequencing Method	Assembly Software
Serosensitive lineage strains								
A9	B743	1.91	27.4	24	220795	41.70	454	Newbler v. 2.6
C031	B620	1.95	31.9	47	122961	41.50	454	Newbler v. 2.6
C10	B621	1.98	28.6	9[?]	412591	41.40	PacBio C1	Celera v. 7.0
F18	B744	1.95	24.3	26	131591	41.70	454	Newbler v. 2.6
F20	B745	1.95	24.4	61	89759	41.60	454	Newbler v. 2.6
F21	B618	1.95	15.7	52	106722	41.50	454	Newbler v. 2.6
F23	B746	1.97	20.5	56	125588	41.50	454	Newbler v. 2.6
F24	B747	1.94	18.9	40	182702	41.70	454	Newbler v. 2.6
N1	B748	1.88	34.1	22	231375	41.60	454	Newbler v. 2.6
N12	B617	1.89	36.1	26	195908	41.60	454	Newbler v. 2.6
N15	B622	1.93	26	45	190441	41.40	454	Newbler v. 2.6
N19	B749	1.94	17.2	71	80324	41.50	454	Newbler v. 2.6
N4	B615	1.94	38.4	46	186739	41.40	454	Newbler v. 2.6
N6	B616	1.95	25.2	40	190206	41.50	454	Newbler v. 2.6
R4	B619	1.95	21.3	41	153844	41.60	454	Newbler v. 2.6
Z18	B750	1.88	17.3	30	126434	41.60	454	Newbler v. 2.6
MX1	212	1.9	449.7	189	16796	41.70	Illumina HiSeq	Newbler v. 2.6
S11	213	1.88	505.5	120	30531	41.50	Illumina HiSeq	Newbler v. 2.6
Serosensitive lineage averages		1.93				41.56		
Seroresistant lineage strains								
BBH18	BBH18	1.86	70	1	Na	41.70	454, Iluminia	Newbler v. 2.6
7169	7169	1.9	182.3	35	176884	41.70	454	Newbler v. 2.3
103P14B1	103P14B1	1.96	33.4	99	74903	41.50	454	Newbler v. 2.0.01.14
12P80B1	12P80B1	1.81	23.3	53	73658	41.70	454	Newbler v. 2.0.01.14
46P47B1	46P47B1	1.85	20.6	69	53508	41.60	454	Newbler v. 2.0.01.14
O35E	B583	1.86	56.9	42	104658	41.70	454	Newbler v. 2.3
BC1	B496	1.95	40.3	43	143587	41.40	454	Newbler v. 2.0.01.14
BC7	B502	1.9	28.7	37	152590	41.50	454	Newbler v. 2.0.01.14
BC8	B503	1.91	43.8	32	173870	41.60	454	Newbler v. 2.0.01.14
C072	B507	1.95	37	25	153832	41.40	454	Newbler v. 2.0.01.14
101P30B1	B508	1.86	33.2	26	215599	41.70	454	Newbler v. 2.0.01.14
RH4	RH4	1.84	200	9	284836	41.60	Illumina	Assembler v. 2.0.0
43617	43617	1.91	NP*	41	89047	41.70	NP*	NP*
Seroresistant lineage averages		1.89				41.60		
Moraxella catarrhalis averages		1.91				41.60		
Divergent strains								
Z7574	XATCC23246	2.23	53.1	26	116763	43.60	Illumina MiSeq	Ray (v2.3.1)
Z7547	FECCUG18198	2.17	13.3	46	87862	43.67	Illumina MiSeq	Ray (v2.3.1)
Z7542	FECCUG18181	2.13	35.9	33	162472	43.69	Illumina MiSeq	Ray (v2.3.1)
Z7546	FECCUG18195	2.12	3.4	196	31433	43.69	Illumina MiSeq	Ray (v2.3.1)
Divergent strains		2.16				43.66		
Moraxella macacae		2.08		1		39.00	454/PacBio/Illumina	Ray (20)

NP* = not provided, sequence obtained online from NCBI.

between the empiric and predicted supragenome sizes was when the divergent strains were added to the combined SS and SR strain lineages, the FSGM predicted 5,373, but the observed was 5,684 indicating that the model is likely underpowered with input from only 4 divergent strains. Combined these data strongly suggest that we have essentially captured the entire supragenome(s) of the species and its component SS and SR lineages, with the probable exception of distributed genes within the divergent group.

We next performed exhaustive pairwise comparisons in which each of the 35 SS, SR, and divergent strains' gene content was compared against every other strains' gene content, resulting in a total of 595 such two-way gene by gene comparisons (supplementary table S4, Supplementary Material online). As expected, the mean gene possession difference values and standard deviations were least within the SR ($\bar{x}$ = 232 ± 55) and SS ($\bar{x}$ = 322 ± 87) clades, with the SR clade showing less divergence than the SS clade. When the two

Table 3

Size of the Core and Supragenomes Based on Strain Number for the three individual *Moraxella catarrhalis* Lineages and the Combined Lineages

	Number of Genes/Genome Type			
Strain_added	**Core**	**Distributed**	**Lineage Supra**	**Species Supra**
Seroresistant lineage strains				
103P14B1	2868	0	2868	2868
12P80B1	2654	266	2920	2920
43617	2616	388	3004	3004
46P47B1	2580	473	3053	3053
7169	2575	488	3063	3063
B496	2574	523	3097	3097
B502	2569	625	3194	3194
B503	2564	644	3208	3208
B507	2563	658	3221	3221
B508	2562	659	3221	3221
B583	2548	675	3223	3223
RH4	2545	690	3235	3235
BBH18	**2541**	**699**	**3240**	**3240**
Serosensitive lineage strains				
B615	2901	0	2901	3457
B616	2737	338	3075	3550
B617	2634	523	3157	3586
B618	2615	590	3205	3615
B619	2603	717	3320	3714
B620	2599	725	3324	3718
B621	2595	753	3348	3729
B622	2595	753	3348	3729
B743	2584	812	3396	3757
B744	2577	847	3424	3767
B745	2572	865	3437	3780
B746	2569	883	3452	3791
B747	2565	898	3463	3801
B748	2560	952	3512	3845
B749	2551	963	3514	3845
B750	2550	964	3514	3845
MX1	2509	1005	3514	3845
S11	**2464**	**1060**	**3524**	**3854**
SR/SS	**2318**	**1536**	**NA**	**3854**
Divergent strains				
Z7542	4106	0	4106	5518
Z7546	3938	168	4274	5518
Z7547	3780	179	4284	5518
Z7574	3632	**180**	4285	**5684**
SS/SR/Divergent	2041	**3643**	NA	**5684**

Note.—Boldface indicates lineage, or lineage combination values.

clades are combined, there was a substantial increase in diversity ($\bar{x}$ = 413 ± 40); but the diversity reached its zenith with the addition of the divergent strains ($\bar{x}$ = 821 ± 809) (table 4). This trend of increasing diversity is also reflected in the standard deviation values for the gene similarity metrics.

Analyses of the various distributed gene frequency classes of the combined SR and SS supragenome revealed small, but easily identifiable, increases in the number of distributed genes that are present in 13 and 18 genomes, relative to the number of distributed genes in the other ($n \neq 13$; $n \neq 18$) intermediate gene frequency classes. These findings correspond with the exact numbers of SR and SS strains in the analysis, and are indicative of the small numbers of genes that are exclusively core to each of these lineages (fig. 2*A*). As expected, a gene-by-gene analysis using our standard clustering criteria (70% homology over 70% of the alignment) (Hogg et al. 2007) showed that in spite of the high degree of similarity between the SR and SS core genomes, there are 33 SR core-specific genes that are not found in any of the SS strains (table 5). Conversely, there are 49 SS lineage core genes that are not found in any of the SR strains. A similar analysis demonstrated a large peak at the gene cluster class size of four when the divergent strains were included in the analysis, indicating that the four divergent strains harbor a unique set of core genes that are not present in the SR and SS lineages (fig. 2*B*). Annotations for all *M. catarrhalis* genes for all of the sequenced strains are available in GenBank (table 1).

Phylogeny and Evolution of the *Moraxella Catarrhalis* Species

We performed three different neighbor-net analyses for the 35 strains (SR, SS, and divergent lineages) using the Splitstree4 program (Huson and Bryant 2006) in which we used as input 1) allelic data for all of the core genes (fig. 3*A*); 2) gene possession data for all of the distributed genes (fig. 3*B*); and 3) a multiple sequence alignment of a concatenation of the LCBs (physical core regions irrespective of gene annotations) as defined by a default MAUVE (Darling et al. 2004) alignment of all 31 *M. catarrhalis* genomes (fig. 3*C*). Finally, we also performed a neighbor-net analysis of the concatenated collinear blocks between all *M. catarrhalis* strains and the *M. macacae* genome (representing an outgroup) (fig. 3*D*). Each of these independent methods demonstrated a clear separation between the SS and SR lineages, but the two SS ribotypes were randomly distributed within the larger SS clade and did not cluster together. It is also apparent that the four divergent strains cluster together, but that they are vastly more different from both the SS and SR lineages than the SS and SR lineages are from each other. All three comparative measures suggest that the divergent strains are slightly closer to the SS strains than the SR strains, but the total distance to the SS strains is >90% of the distance to the SR strains. The reticulations (webbing) present within these diagrams are indicative of ambiguous ancestry, but given that *M. catarrhalis* is naturally highly competent and transformable these findings are suggestive of HGT events as they represent alternative phylogenetic branching; however, it is also possible that they represent parallel mutations, or the loss of ancestral copies in one or the other lineages These reticulations are present both within the individual

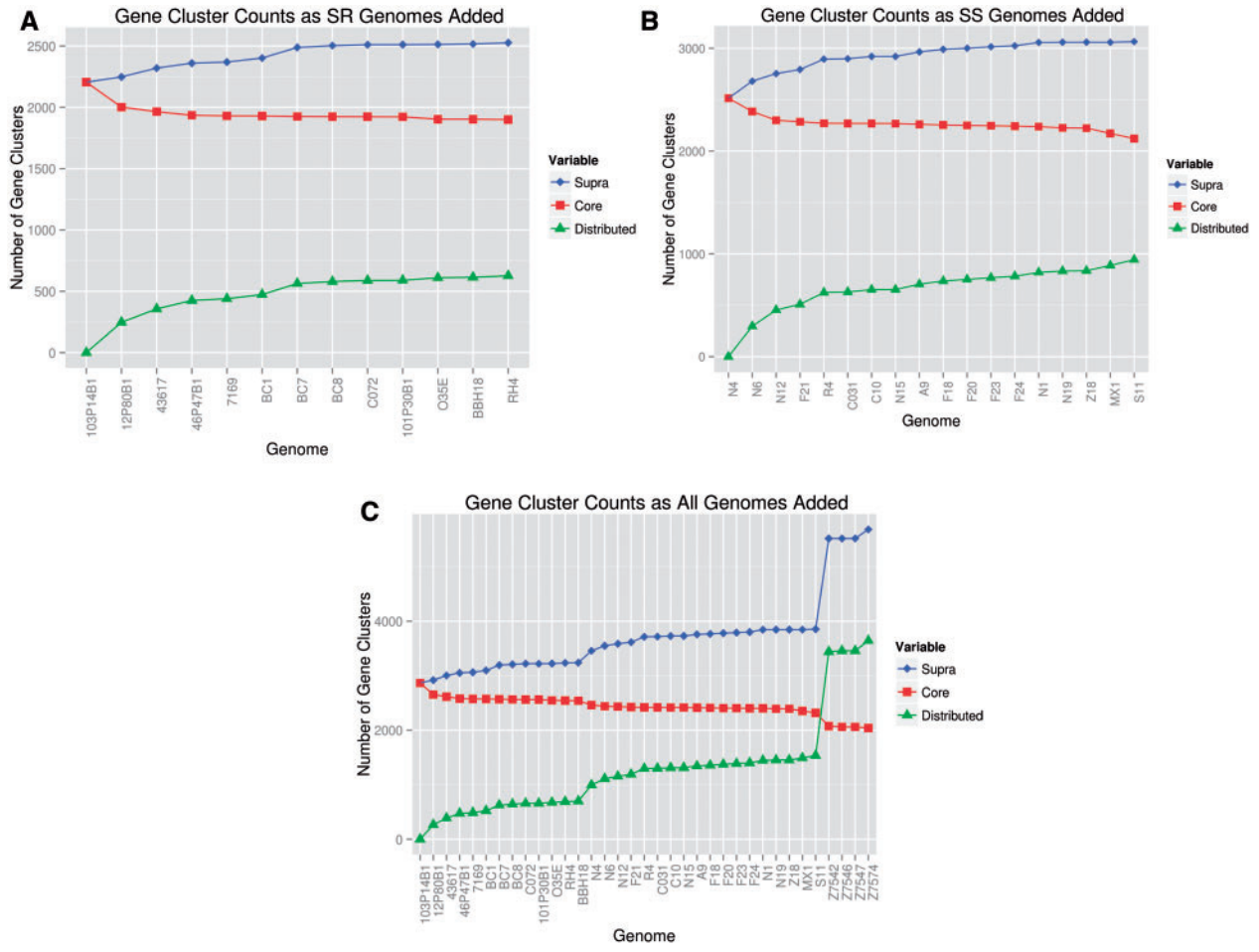

Fig. 1.—*Moraxella catarrhalis* lineage and species supragenome, core genome, and distributed genome plots. *Y*-axis = number of gene clusters; *x*-axis shows the name of each strain added to the analysis. (*A* and *B*) SR and SS lineages respectively: The blue line and diamonds show the size of the supragenome as each individual strain's genome is added; the red line and squares show the size of the core genome as each strain is added; and the green line and triangles show the increase in the size of the distributed genome as each strain is added. (*C*) Species level plots of the supragenome, core genome, and distributed genome as individual strains are added to the analysis. The seroresistant strains were added first (strains 1–13), then the serosensitive strains (strains 14–31), and finally the divergent strains (strains 32–25). Note the distinct increase in the size of the supragenome associated with the addition of the first several serosensitive lineage strains indicative of their larger distributed genome with respect to the SR lineage. In contrast, the core genome is largely unaffected by the addition of the SS lineage strains. Also note the extraordinary increases in the sizes of the supra- and distributed genomes when the divergent strains are added, and the large decrease in the size of the core genome.

lineages, and are also present between the lineages. Such findings are consistent with interlineage gene flow, but could also indicate parallel mutations or loss of ancestral genes. In particular, it is clear that there is substantial distributed gene flow between the SR and SS lineages, indicating that they are not currently reproductively isolated. The decreased number of reticulations in the concatenated core region tree is a direct result of the selection process for inclusion of regions in the analysis as only regions that are core across all strains were included. The concatenation network in which we included *M. macacae* as an out group showed that it was approximately equidistant from all three lineages.

We then evaluated the various groupings for evidence of diversity generation by determining both the intra- and interlineage levels of HGT, and by determining the extent of allelic diversity as measured by SNPs of core genes within and between each group as a measure of their relative phylogenetic ages. These last two measures were designed to determine if there were different evolutionary mechanisms acting on the two main lineages as has been previously suggested (Wirth et al. 2007). These investigators speculated that the SR and SS lineages evolved primarily by different mechanisms (Wirth et al. 2007). However, we find no support for this conjecture as it is clear that HGT of distributed genes is just as prevalent among the SS lineage as the SR lineage, and that

Table 4

Pairwise Genomic Comparison Statistics among *Moraxella catarrhalis* Strains

Measures		SR	SS	SR and SS	SR, SS, and Div
Similarity	Min	2,595	2,681	2,484	2,254
	Max	2,756	2,897	2,897	4,279
	Average	2,668	2,754	2,637	2,591
	Standard deviation	35	52	82	203
Difference	Min	107	4	4	4
	Max	351	445	671	2,531
	Average	232	322	413	821
	Standard deviation	55	87	140	809
Comparison	Min	2,298	2,185	1,866	−222
	Max	2,611	2,893	2,893	4,273
	Average	2,436	2,402	2,224	1,770
	Standard deviation	68	141	211	954
PairUnique	Min	0	0	0	0
	Max	25	16	25	25
	Average	1	<1	<1	0.17
	Standard deviation	3	1.4	1.5	1.35

NOTE.—Similarity = the number of gene clusters shared between strain pairs; difference = the number of gene clusters not shared between strain pairs; comparisons = the difference of the similiarity and difference score; PairUnique = the number of gene clusters shared only by a single pair of strains; SR = seroresistant lineage; SS = serosenstive lineage; Div = divergent strains; all numbers refer to the number of gene clusters.

exchange of different core gene alleles between and among strains of all lineages is also consistent, the only reason that more allelic variation is observed in the SS lineage is due to the age of the lineage.

Next we performed an analysis of the core genes from all of the strains to look for regions across the *M. catarrhalis* chromosome evidencing the highest incidences of HGT using the orderedPainting software (Yahara et al. 2014). This tool was specifically designed to perform HGT-based population genomic studies of prokaryotes. Analyzing the genome as a whole it can be can be seen that the rates of recombination vary greatly by locus (fig. 4*A*). Chromosomal regions with intensity factors greater than 236 are in the top one percentile with respect to recombination rates. Examination of figure 4*B* shows that the loci experiencing high rates of HGT are not clustered together, but are instead found throughout the genome; however, there are clearly genomic regions that experience highly divergent rates of recombination. The *M. catarrhalis* core locus with the highest recombination rate was identified as an operon containing two amidotransferase subunit genes of unknown specificity. Interestingly, in more than two-thirds of all the SR/SS genomes examined, a ß-lactamase gene (*bro*) (fig. 5), responsible for penicillin resistance, was inserted between the two amidotransferase genes, A and B. This distributed gene which has two alleles (*bro*1 and *bro*2), either of which were found inserted between the amidotransferase genes of a given strain, was found in all but one of the SR lineage strains, and in slightly more than half of the SS strains; it was not identified in any of the divergent strains. Independent phylogenetic trees were built using phyml from mauve alignments containing the amidotransferase A and *bro* genes, respectively (fig. 6*A* and *B*). Examination of the amidotransferase A tree establishes that this gene is frequently moving via HGT both within each of the SR and SS clades, but also between the clades as well. We see this gene tree structure, which is very different from a consensus tree for the species as a whole, presumably because of the positive selective pressure produced by acquisition of the adjacent *bro* gene which provides for high-level penicillin resistance. Examination of the data suggests that the SS strains F18 and MX1 acquired their amidotransferase and *bro*1 genes via HGT from an SR strain within/related to the BC7-containing group of SR strains. Similarly, it is likely that the O35E and BC1 SR strains acquired their amidotransferase A and *bro*1 genes via HGT from an SS strain within/related to the F-21/F24 containing group. Within the BC7 group of SR strains there has also been loss of the *bro1* gene (46P46B1) followed by acquisition of the *bro*2 allele in several strains (BBH18, 12P80B1, and 43617). The ß-lactamase gene itself was not evaluated in the orderedPainting analysis because it is a distributed gene, albeit a highly prevalent one.

The independent evolution of the SS and SR lineages into distinct clades with neither appearing to be ancestral to the other raises two questions: When did the split occur that resulted in their isolation, and when were they reunited? Toward this end we attempted to perform an evolutionary divergence analysis with BEAST (Drummond and Rambaut 2007); however, due to our having only a single unverifiable calibration point extreme care must be used in interpreting these data. The BEAST results suggest an earlier split between the *M. macacae* and *M. catarrhalis* populations at 37 Ma with an effective sample size (ESS) score of 118, and a split between the SS and SR *M. catarrhalis* clades at 0.265 Ma with an ESS score of 3 after 10 million iterations. It is important to note that an ESS score of less than 100 is considered not significant; however, without intermediate time points (i.e., additional phylogenetic branches within *Moraxella*) it is not possible to further refine this estimate.

Virulence Factor Distribution

To date, numerous *M. catarrhalis* virulence factors have been described that play a role in immune evasion, cellular adherence, biofilm formation, and survival during iron starvation, which have been reviewed (de Vries et al. 2009; Su et al. 2012). Previous evaluation of virulence factor distribution in 12 SR lineage isolates (Davie et al. 2011) demonstrated that, with a few exceptions, all known virulence factors were

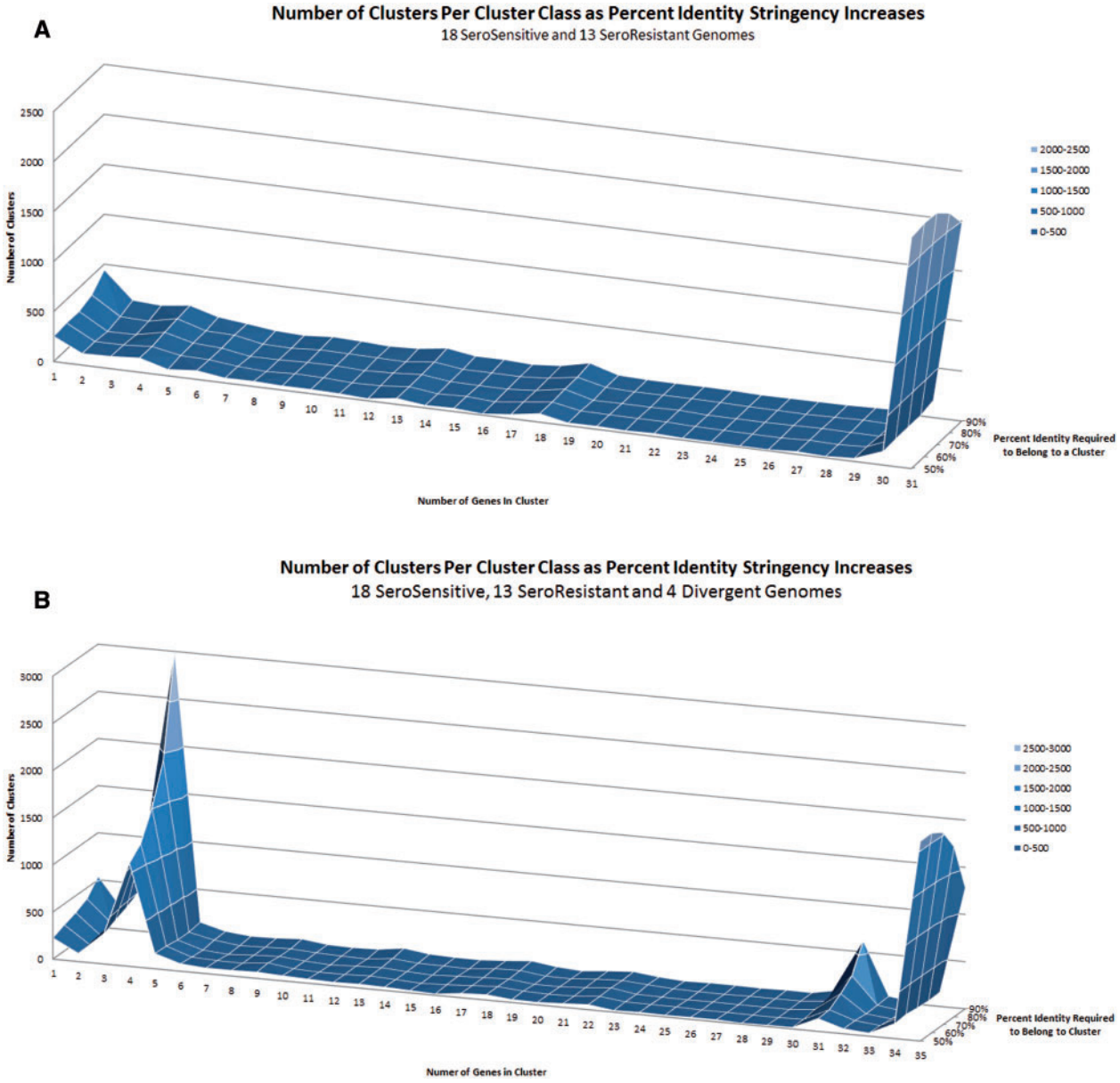

Fig. 2.—Number and distribution of gene clusters as a function of frequency within the species, and as the percent similarity requirement increases. *x*-axis = number of genomes containing a given gene cluster; *y*-axis = number of gene clusters; *z*-axis = percent identity to be grouped in a cluster. (*A*) Analysis of the combined SR and SS lineages. Note the small increases in the gene frequency classes at 13 and 18 which correspond to the numbers of SR and SS strains analyzed; these represent genes which are uniquely core to each lineage; (*B*) analysis of the combined SR, SS, and divergent lineages. Note the very large increase in the gene frequency class at four genomes corresponding to number of divergent strains in the analysis; this spike represents genes that are core to the divergent lineage which are not present in either the SR or SS lineages.

present in each SR lineage strain. In this study, we found 33 gene clusters that were core to the SR lineage and missing from all of the SS strains (table 5). None of these gene clusters are obvious virulence factors; however, the majority of these clusters that are unique to the SR lineage either are predicted to encode hypothetical proteins of unknown function or are under-annotated, suggesting that there remains much important biology to be developed to understand the observed phenotypic differences. Within the small group of SR-specific genes, almost all of those that were annotated were associated with phosphate metabolism, including at least ten genes with putative annotations as phosphodiesterases, phosphatases, phosphate permeases, and adenosine triphosphate (ATP) transporters clustered together on the chromosome and likely assembled into one or more tightly linked operons. In addition, there were also a number of orphan

Table 5

Seroresistant Lineage Core Genes Not Present among the Serosensitive Lineage Strains

BBH18_Genes	BBH18_annotations
MCR_0971	Hypothetical protein
MCR_0970	Hypothetical protein
MCR_0969	Hypothetical protein
MCR_0968	Hypothetical protein
MCR_0646	IS200 family transposase
MCR_0640	CRISPR-associated protein NE0113 (Cas_NE0113); family protein
MCR_0623	Acid phosphatase autotransporter
MCR_0619	Hypothetical protein
MCR_1427	Hypothetical protein
MCR_0008	Glycerophosphoryl diester phosphodiesterase
MCR_1871	Putative membrane protein
MCR_1721	HAD-superfamily subfamily IB hydrolase
MCR_1723	Polyamine ABC transporter permease protein
MCR_1725	Type I phosphodiesterase/nucleotide; pyrophosphatase
MCR_1726	Extracellular solute-binding protein family 1; protein
MCR_1731	Phosphate ABC transporter ATPase subunit PstB
MCR_1732	Phosphate ABC transporter permease protein PstA
MCR_1733	Phosphate ABC transporter permease protein PstC
MCR_1734	Phosphate ABC transporter substrate binding; protein PstS
MCR_0920	Conserved hypothetical protein
MCR_1199	Hypothetical protein
MCR_1722	Polyamine ABC transporter ATPase subunit
MCR_1724	Polyamine ABC transporter permease protein
MCR_1013	Hypothetical protein
MCR_1123	Conserved hypothetical protein
MCR_0622	Hypothetical protein
MCR_0811	Hypothetical protein
Unannotated BBH18 Genes	Annotations
103P14B1_285	Hypothetical protein
12P80B1_301	Serine protease
12P80B1_1567	Hypothetical protein
B496_1518	Hpothetical protein
B583_283	Hypothetical protein

Note.—BBH18 genes refer to SR core genes that were annotated in the SR strain BBH18 [13]. This genome was chosen as it has been closed. Unannotated BBH18 genes refer to SR core genes that were annotated in one of the other SR strains, but not in BBH18; however, all SR strains including BBH18 were determined by FASTA alignment [16,44] to contain these genes.

phosphatases, phosphodiesterases, and transporter genes associated with phosphate flux

The known virulence factor genes were found also, somewhat surprisingly, to be nearly universally conserved across the panel of sequenced SS strains (supplementary table S5, Supplementary Material online). One of the few exceptions was the gene encoding the Opa-like protein A (*OlpA*) which was absent from strain ATCC 43617, confirming previous findings (Brooks et al. 2007). The prevalence of genes encoding the *M. catarrhalis* filamentous hemagglutinin-like proteins (*Mha/Mch*) (Balder et al. 2007; Plamondon et al. 2007) could not be accurately analyzed due to high sequence similarity and inconsistent assembly of whole-genome sequence data, as previously pointed out (Davie et al. 2011). All of the strains contained the *uspA2/A2H* gene. The *uspA1* gene was detected in all SR isolates and in 16 of the 18 SS isolates, with isolates N4 and N15 being *uspA1* negative, confirming previous PCR-screening results by Bootsma et al. (1996). These results are comparable with those reported in a publication by Verhaegh *et al.* (2011) who showed that in a panel of 112 *M. catarrhalis* isolates (n = 22 SR lineage, n = 90 SS lineage based on ribotyping), 97% were *uspA1* PCR positive, and all strains were PCR positive for either *uspA2* or *uspA2H.* In that study, *mid/hag* was only detected by PCR in 80% of isolates, while it was present in 100% of the isolates in our study.

Interestingly, the presence of the β-lactamase gene *bro-1/2* differed between the lineages as out of the 18 sequenced SS strains, 8 did not harbor the β-lactamase gene, in contrast to only two *bro-1/2* negative SR isolates. In silico typing of LOS gene clusters (Edwards et al. 2005) demonstrated that most isolates harbor the type A gene cluster (24/31), although the prevalence was higher in the sequenced SS lineage isolates (17/18) than in the SR lineage (7/13), which is in line with previous observations (Verhaegh et al. 2008). The presence of *lgt4*, encoding a predicted N-acetylglucosylamine transferase, is characteristic for LOS type A and C strains (Peak et al. 2007). In our panel of sequenced isolates, an LOS type B gene cluster lacking *lgt4* was only found in four SR lineage isolates and not in any SS lineage isolates. Furthermore, the LOS type C gene cluster was detected in 1 and 2 SS and SR lineage isolates, respectively.

Phenotype–Genotype Association

To enable phenotype–genotype association analyses for key virulence determinants in *M. catarrhalis*, that is, complement resistance and cellular adherence, the SS and SR strains used in these analyses were subjected to phenotyping with respect to their ability to survive in 40% NHS and adherence to respiratory tract epithelial cells. These studies demonstrated that all but one (ATCC 43617) of the strains previously determined by gene-based studies (Davie et al. 2011) to belong to the SR group were intermediate to highly resistant to NHS, whereas all of the strains previously characterized as SS were highly sensitive to killing by NHS (fig. 7*A*). Similarly, all but two (C072 and 7169) of the SR lineage strains showed efficient adherence to A549 type II alveolar cells and Detroit 562 pharyngeal epithelial cells, whereas all but one (N1) of the SS strains adhered poorly to these human cells types (fig. 7*B*). All these phenotypic findings were statistically significant. Next we analyzed the association of *M. catarrhalis* virulence phenotypes (fig. 7*A* and *B*) with genotypes using PhenoLink

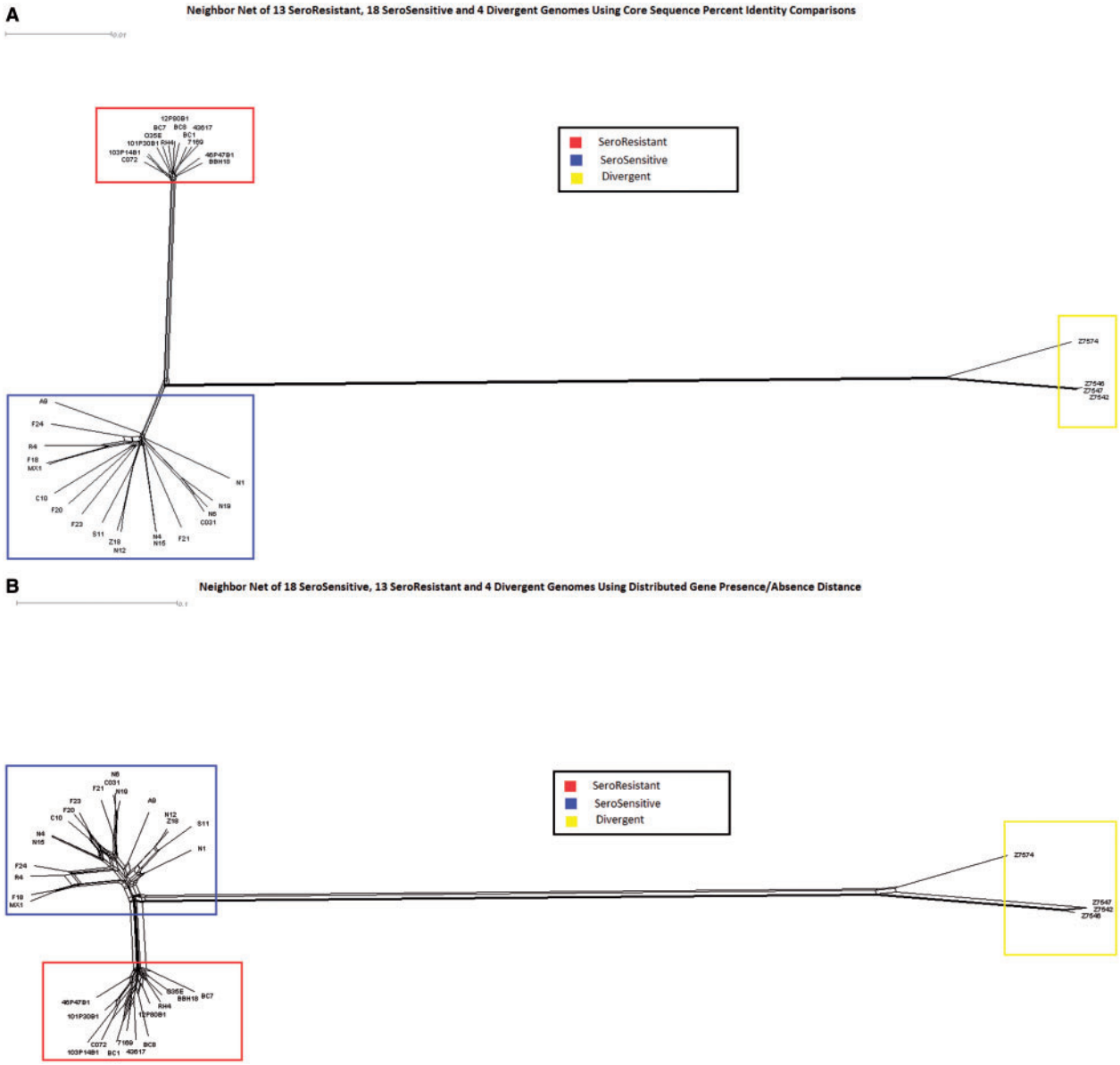

Fig. 3.—NeighborNet analyses performed using SplitsTree4. Branch length is proportional to the degree of variation between and among strains. Reticulations (webbing) are indicative of HGT as they represent ambiguities with respect to lineal descent. Blue boxes indicate SS strains; red boxes indicate SR strains; and yellow boxes indicate divergent strains. (*A*) Network prepared using core gene allelic data; (*B*) network prepared using presence/absence of distributed gene clusters; (*C*) network prepared using concatenated core sequences independent of annotation; (*D*) network prepared using core sequence alignment shows phylogenetic distance between *Moraxella catarrhalis* and *Moraxella macacae*. The latter was used as outgroup to root the tree and to look for any ancestral relationships within the *M. catarrhalis* species. Note that the greatest degree of HGT both within and between lineages is observed with the distributed gene analyses; however, it is also evident that the core genes and regions are exchanged as well.

(Bayjanov et al. 2012). However, given the clear separation of the SR and SS lineages and the ability of the SR stains to resist complement-mediated killing, the value of this analysis was limited, yielding primarily the lineage-specific genes identified above. Moreover, manual comparison of the SR lineage isolates with strain ATCC 43617, the only complement-sensitive SR lineage isolate in our panel, showed that there are no genes that were absent only from the ATCC 43617 genome, suggesting this defect is allelic in nature. Similar observations were made for association with A549 adherence phenotypes, as all SS lineage isolates were found to attach to A549 cells with low efficiency. Comparing the low binding SR isolates (7169 and C072) with the intermediate and/or high binding SR isolates did not reveal any genes that were lacking specifically from these two isolates. Association analyses for the Detroit 562 adherence phenotype showed that two

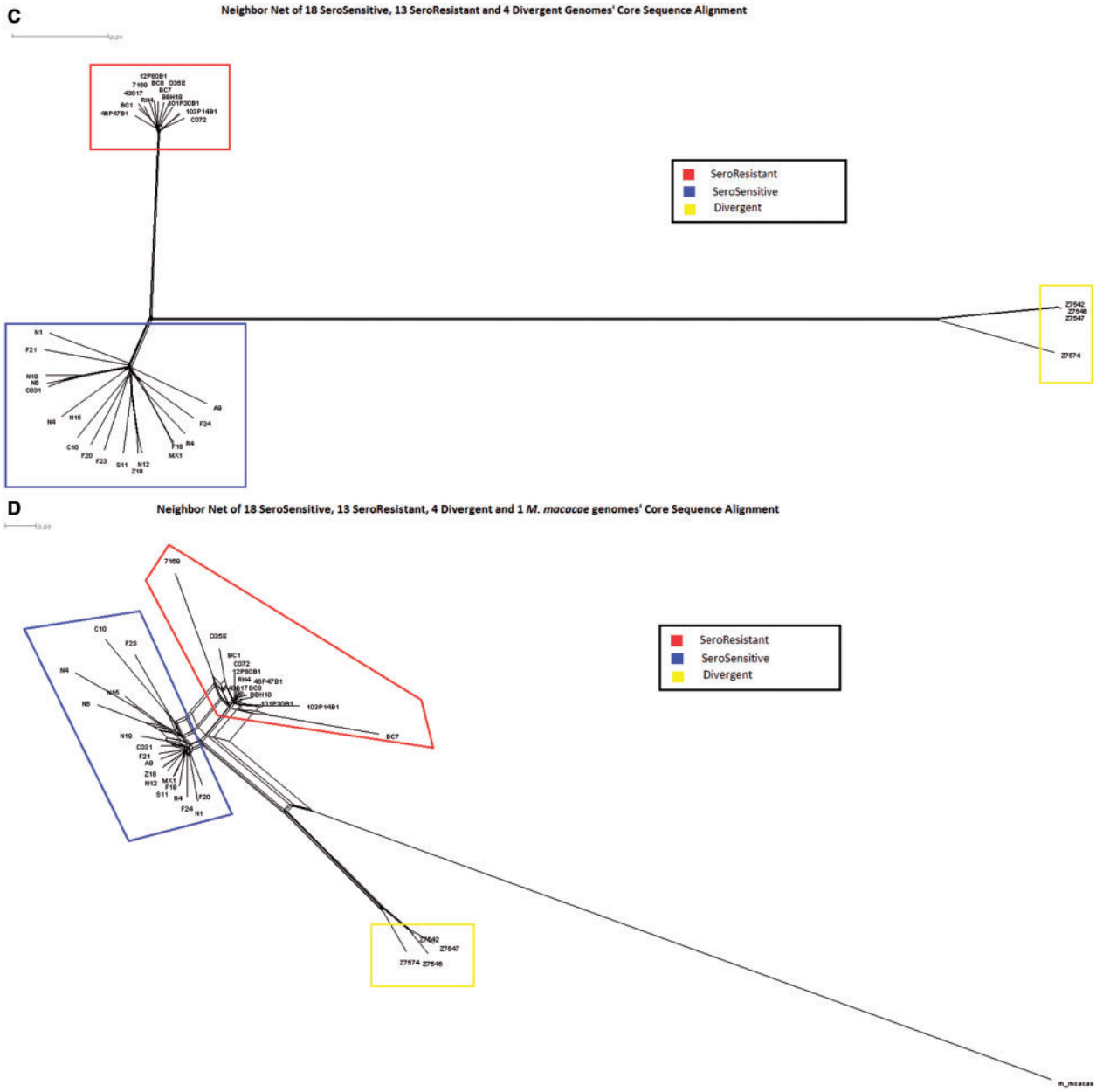

Fig. 3.—Continued.

genes (*MCR_1831* and *MCR_1286*, encoding a type I restriction modification DNA specificity protein and a membrane protein-like protein, respectively) were present in the high binding SS isolate N1 and absent from all other SS lineage isolates; however, these genes were also found to be present in the low binding SR lineage isolates 103P14B1, 12P80B1, and BC7.

Discussion

The relative overall average pairwise gene possession differences among the SR lineage strains (Davie et al. 2011) is only 232 ± 55, which is substantially less than has been observed for two other supragenomically characterized nasopharyngeal pathogens of similar genomic size—*H. influenzae* (395 ± 4) (Hogg et al. 2007) and *S. pneumoniae* (407 ± 91) (Hiller et al. 2007). With the inclusion of the SS lineage strains in this study, the species-level pairwise gene possession differences (412 ± 140) very nearly approximate those of the NTHi and pneumococcus. The somewhat larger SD value is attributable to the *M. catarrhalis* species being composed of two distinct lineages such that all cross-lineage pairwise comparisons result in very much higher numbers of differences than do any of the within lineage comparisons.

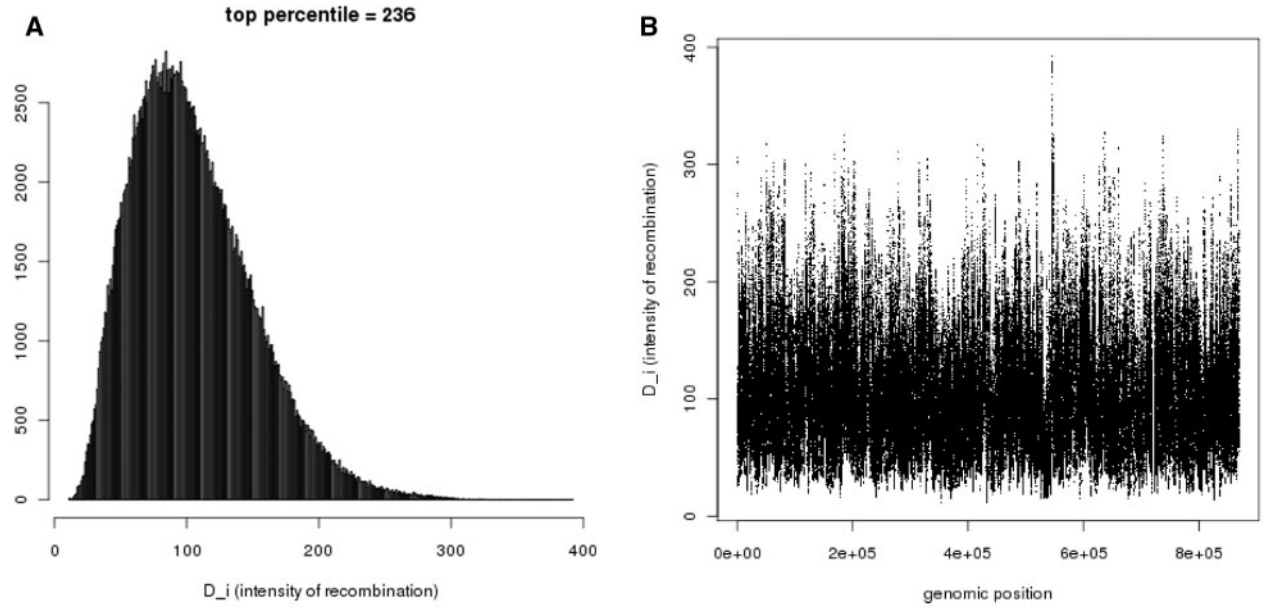

Fig. 4.—(A) Output from the orderedPainting program showing the number of SNPs on the *y*-axis and the intensity of the recombination signal (evidence of HGT) on the *x*-axis. SNPs in the top percentile had *D_i* values of ≥ 236. (*B*) Graph of the *D_i* values across the *Moraxella catarrhalis* chromosome. The locus with the highest intensity (~400) corresponded to the aminotransferase A gene.

Moreover, the supragenomes for the SR, SS, and combined SS/SR *M. catarrhalis* groupings based on the FSM predictions are very close in size. This is an indication that there is very little divergence between the two lineages and is highly supportive of a single species model. If we had observed a dramatic decrease in the size of the *M. catarrhalis* SR core genome upon the addition of the SS lineage strains, as was seen with the combining of multiple *Gardnerella vaginalis* clades (Ahmed et al. 2012), that would have been supportive for a species-level taxonomic split. In spite of the relative similarity between the SR and SS clades, they each possess a small clade-specific core genome. These genes are, respectively, excellent candidates for virulence analyses and suppressors of pathogenicity. It is tempting to hypothesize that the cluster of genes associated with phosphate metabolism (table 5) that are core to the SR lineage may be involved in a nucleotide second messenger system that controls virulence gene expression (D'Argenio and Miller 2004). The fact that many traditional virulence genes were highly conserved across the SS and SR lineages indicates that individual gene presence is not predictive of virulence, which suggests that it may be gene combinations, differences in expression, or environmental conditions that in concert control differences with regard to virulence potential. In support of this last hypothesis it has been shown that host factors play a role in determining whether/when given *M. catarrhalis* strains are pathogenic or not as previous studies have shown that under proinflammatory conditions that host tissues upregulate CEACAMS which then provide for better binding by SS strains (Hill and Virji 2003; Hill et al. 2012).

The vast majority of *M. catarrhalis* strains that are recovered from respiratory tract infections belong to the SR lineage and it has been noted for the past quarter century that almost all such isolates are penicillin resistant due to the possession of a ß-lactamase gene (*bro*) (Bluestone et al. 1992; Johnson et al. 2003). Using an unbiased screen for identifying core chromosomal sites undergoing high rates of HGT, we identified an amidotransferase operon. Examination of this locus revealed that it is the sole chromosomal site for insertion of the distributed *bro* gene for both the SR and SS lineages. The high rate of recombination at this site combined with the very limited heterogeneity of the *bro* genes, in contrast to the high degree of allelic variability among the adjacent aminotransferase A genes, is supportive of a recent acquisition history for the *bro* gene followed by rapid spread throughout the *M. catarrhalis* species confirming earlier findings (Bootsma et al. 1996, 1999, 2000b).

In the current analyses, we also included a group ($n = 4$) of highly divergent strains that had previously been characterized as distantly related to *M. catarrhalis* (Wirth et al. 2007). The inclusion of these strains essentially doubled the average number of pairwise gene possession differences (821 ± 809) when compared with the SR/SS species-level view. What was also striking about the supragenome comparisons when these divergent strains were added was that we obtained some negative pairwise comparison scores (table 4) which means that some strains shared less than 50% of their genes. In our analyses of the supragenomes of some two dozen bacterial species, the only other time that we obtained negative-value

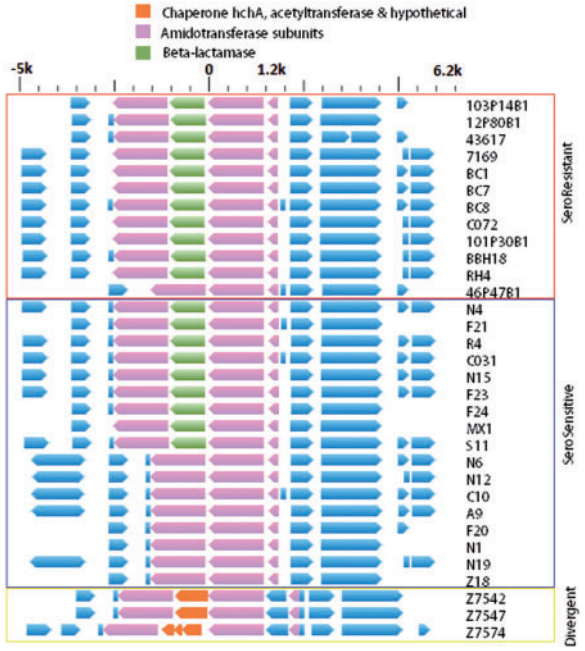

Fig. 5.—Chromosomal locus within the *Moraxella catarrhalis* genome corresponding to the region of greatest recombination frequency among all strains identified by orderedPainting analysis. The gene composition for this locus for each of the sequenced strains (excepting one SR strain, one SS strain, and one divergent where contig breaks occurred in this locus) is listed with the strain designation on the right. The pink arrows refer to the core genes encoding the two amidotransferase subunits; the green arrows indicates the distributed ß-lactamase gene (*bro*) inserted into the SR and SS strains; the orange arrows refer to the genes (no homology to the ß-lactamase gene) inserted between the amidotransferase genes in the divergent strains; and the blue arrows refer to flanking genes outside of the amidotransferase operon. Flanking regions of 5-kb upstream of start codon, and downstream of stop codon of this gene. The scale bar at the top is in kilobases.

comparison scores was with *Gardnerella vaginalis* (Ahmed et al. 2012) which turned out to actually be a genus-level classification composed of multiple species.

As with the *Gardnerella* work we now propose that the divergent strains belong in a separate taxonomic grouping. Although the definition of a bacterial species is currently under critical review by the scientific community, our suggestion is supported by multiple types of analyses and metrics that are both broadly employed, and currently viewed as being the current "best estimate" used to define a split between different strains and different species. The currently accepted level of similarity to be included in the same species with regard to nucleotide identity is an ANI of 95–96%, and a corresponding tetranucleotide identity of >99% (Konstantinidis and Tiedje 2005; Goris et al. 2007; Richter and Rosello-Mora 2009; Chan et al. 2012). Using these criteria, these data are supportive of the SR and SS strains belonging to the same species. However, the average pairwise comparisons between SS/SR strains and the divergent strains gave ANI values of 89.01% and tetranucleotide values of 95.57%, both far outside the accepted range for a species.

Similarly, each species has a very narrowly defined percent GC content which always varies by less than 1%, and usually by less than 0.5%, among the component strains. All of the SS and SR lineage strains have GC contents within 0.3% of one another, again supporting their classification as a single species. As with the ANI values, the four divergent strains cluster together very closely among themselves, but their GC content of 43.66% differs by more than 2% from the SR/SS lineages, again clearly indicating that they are *sui generis.* Finally, analysis of the effect on the core genome size by combining the various lineages is also very revealing with regard to identifying natural genomic proclivities. Riley and Lizotte-Waniewski (2009), Donati et al. (2010) have proposed using the core genome as a test of whether individual strains should be grouped within the same species; they showed that even a single *Streptococcus mitis* strain's genome added to the *S. pneumoniae* supragenome resulted in a precipitous drop in the combined core genome size. Thus, we have proposed a bacterial species definition based on core genome relatedness such that any strain that results in less than a 10% decrease in the size of an established (at or near its asymptote) core genome belongs within that species and conversely those that result in a greater than 10% decrease belong in a different species (Nistico et al. 2014). Although many of the pathogenic enterobacteriaceae species such as *Escherichia coli* have very large supragenomes (Ahmed et al. 2012), comparison across clades shows little variability among core genomes. In contrast, the addition of just a single member of the *M. catarrhalis* divergent strain group to the combined SS/SR core genome results in decrease in size of greater than 11%. Collectively, these data are supportive of the current taxonomy in which the SR and SS lineages form a single *M. catarrhalis* species, but suggest that the four divergent strains may be considered a distinct genospecies.

The finding that the *M. macacae* is approximately equidistantly related to the two (SR and SS) *M. catarrhalis* lineages is suggestive that neither of the lineages are directly ancestral to the other. However, the independent evolution of the SS and SR lineages raises two questions: When did the split occur that resulted in their isolation, and when were they reunited? Based on their possession of nearly identical core genomes and the amount of horizontal gene flow between the two lineages, they are clearly still a single species by any criteria; however, the low-resolution BEAST analysis we performed suggested a split at ~265,000 BP. It is interesting to speculate that one of the lineages may have evolved independently in the Neanderthal population which diverged from the ancestors of modern humans ~ 370 Ma (Noonan et al. 2006) and was then subsequently reintroduced into anatomically modern humans, where the other lineage evolved, during the short period of geographic and social overlap of these

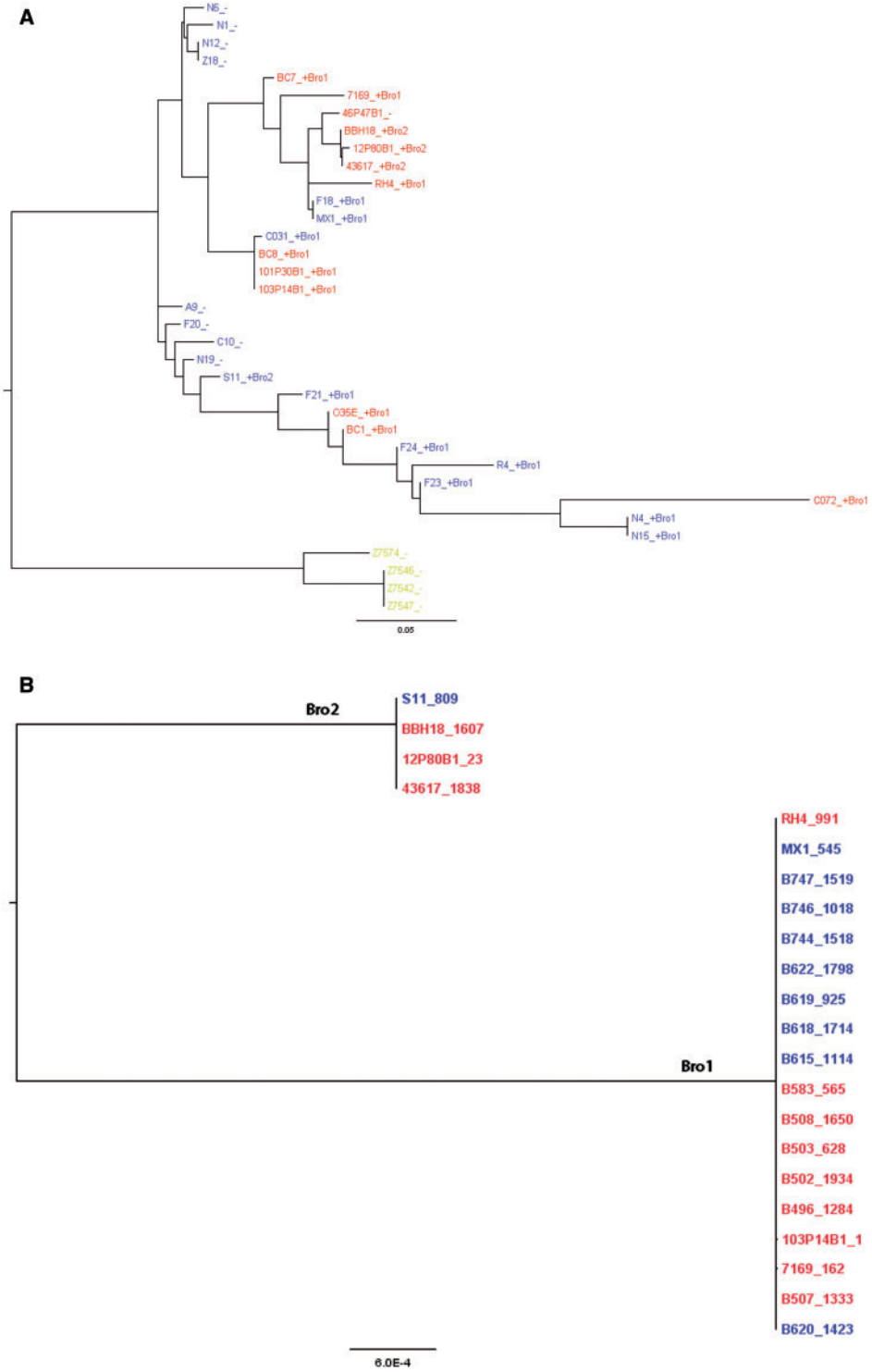

Fig. 6.—(*A*) Phylogenetic tree built using phyml from a mauve alignment of the amidotransferase subunit A genes. SR strains are shown in red typeface; SS strains are shown in blue typeface; and the divergent strains are shown in yellow typeface. "+" indicates the presence of the ß-lactamase gene in the strain, and "-" indicates a lack of the ß-lactamase gene in the tree. Bro1 and Bro2 following the + after the strain name refers to which of the two ß-lactamase alleles are present in the strain; and (*B*) phylogenetic tree built using phyml from a mauve alignment of the ß-lactamase gene (*bro*); note that there are only two alleles of the *bro* gene distinguished by five-point mutations of which all but one are synonymous. The *bro*2 containing strains are on the left, and the *bro*1 containing strains are on the right. SR strains are shown in red typeface; SS strains are shown in blue typeface.

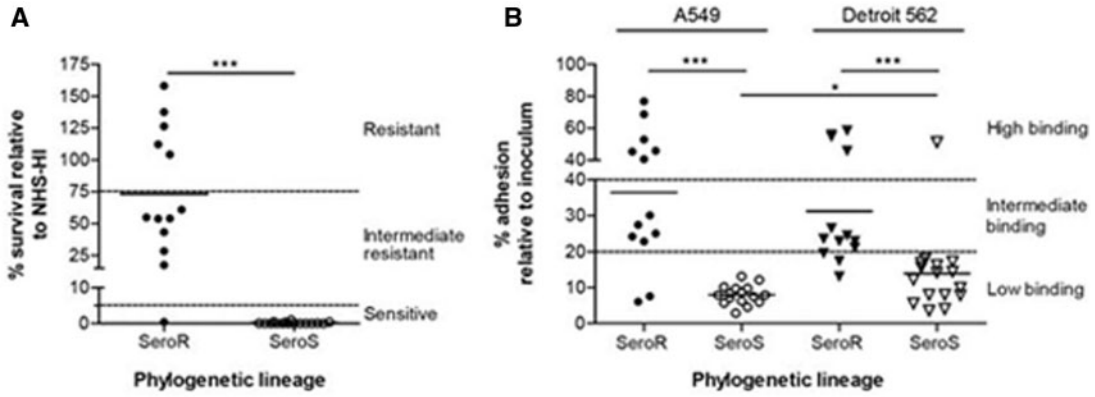

Fig. 7.—Phenotypic characterization of sequenced *Moraxella catarrhalis* strains. (*A*) Survival in 40% NHS. Strains of the SR lineage were intermediate to highly resistant to the action of NHS, except strain ATCC 43617, whereas all strains of the SS lineage were efficiently killed in the presence of active complement. Percentage survival ($n \geq 3$) in NHS is expressed relative to the survival in NHS-HI. (*B*) Adhesion characteristics of *M. catarrhalis* strains to A549 type II alveolar and Detroit 562 pharyngeal epithelial cells. SS lineage strains showed less efficient binding to both respiratory tract epithelial cell lines as compared with SR lineage isolates, which attached to both cell lines with variable efficiency. Binding of SS lineage isolates was more efficient to pharyngeal epithelial cells than to alveolar epithelial cells. Adherence is expressed relative to the inoculum ($n \geq 4$). Statistical difference (*A* and *B*) was determined with a Mann–Whitney test with $*P < 0.05$, $**P < 0.01$, $*P < 0.001$.

distinct people (Green et al. 2010) in Europe approximately 30–45,000 years before the present following the latter's migration out of Africa. Although the estimated separation dates are somewhat different, they are still supportive of the general hypothesis of short-term independent lineage evolution as it is very difficult to perfectly synchronize evolutionary clock rates across species particularly those as divergent as microbe and man. However, these data could also support other early human population splits and reunions.

Summary

The analyses presented herein support the hypothesis that the *M. catarrhalis* species consists of two distinct lineages, the SR and the SS, with separate evolutionary histories and small, but distinct, lineage-specific core genomes; however, it is also clear that the two clades have been reunited and now share a single environmental niche and regularly exchange DNA via HGT. Thus, the questions of how and when the lineages diverged and then at some later point reconverged need to be further addressed. It is also clear that the divergent strains form a distinct genospecies separate from *M. catarrhalis*.

Supplementary Material

Supplementary tables S1–S5 and figure S1 are available at *Genome Biology and Evolution online* (http://www.gbe.oxfordjournals.org/).

Acknowledgments

This work was supported by Allegheny Singer Research Institute, Allegheny General Hospital, and Drexel University College of Medicine with grant funding from the United States National Institutes of Health grants DC05659, AI080935, and DC02148, to G.D.E., and from a Vienna Spot of Excellence grant (ID337956). The authors thank Mark Achtman for providing access to the DNA for the divergent strains and for facilitating their whole-genome sequencing; Joshua Chang Mell for fruitful discussions; and Ms Mary O'Toole and Ms Carol Hope for help in the preparation and submission of this manuscript.

Literature Cited

Ahmed A, et al. 2012. Comparative genomic analyses of seventeen clinical isolates of *Gardnerella vaginalis* provide evidence of multiple genetically isolated clades consistent with subspeciation into genovars. J Bacteriol. 194:3922–3937.

Aul JJ, et al. 1998. A comparative evaluation of culture and PCR for the detection and determination of persistence of bacterial strains and DNAs in the *Chinchilla laniger* model of otitis media. Ann Otol Rhinol Laryngol. 107:508–513.

Aziz RK, et al. 2008. The RAST Server: rapid annotations using subsystems technology. BMC Genomics 9:75.

Balder R, Hassel J, Lipski S, Lafontaine ER. 2007. *Moraxella catarrhalis* strain O35E expresses two filamentous hemagglutinin-like proteins that mediate adherence to human epithelial cells. Infect Immun. 75:2765–2775.

Bayjanov JR, Molenaar D, Tzeneva V, Siezen RJ, van Hijum SA. 2012. PhenoLink—a web-tool for linking phenotype to ~omics data for bacteria: application to gene-trait matching for *Lactobacillus plantarum* strains. BMC Genomics 13:170.

Bluestone CD, Stephenson JS, Martin LM. 1992. Ten-year review of otitis media pathogens. Pediatr Infect Dis. 11(8 Suppl):S7–S11.

Boissy R, et al. 2011. Comparative supragenomic analyses among the pathogens *Staphylococcus aureus*, *Streptococcus pneumoniae*, and *Haemophilus influenzae* using a modification of the finite supragenome model. BMC Genomics 12:187.

Bootsma HJ, van der Heide HG, van de Pas S, Schouls LM, Mooi FR. 2000a. Analysis of *Moraxella catarrhalis* by DNA typing: evidence for a distinct subpopulation associated with virulence traits. J Infect Dis. 181:1376–1387.

Bootsma HJ, van Dijk H, Vauterin P, Verhoef J, Mooi FR. 2000b. Genesis of BRO beta-lactamase-producing *Moraxella catarrhalis*: evidence for transformation-mediated horizontal transfer. Mol Microbiol. 36:93–104.

Bootsma HJ, van Dijk H, Verhoef J, Fleer A, Mooi FR. 1996. Molecular characterization of the BRO beta-lactamase of *Moraxella (Branhamella) catarrhalis*. Antimicrob Agents Chemother. 40:966–972.

Bootsma HJ, et al. 1999. *Moraxella (Branhamella) catarrhalis* BRO betalactamase: a lipoprotein of gram-positive origin? J Bacteriol. 181:5090–5093.

Borriello G, Richards L, Ehrlich GD, Stewart PS. 2006. Arginine or nitrate enhances antibiotic susceptibility of *Pseudomonas aeruginosa* in biofilms. Antimicrob Agents Chemother. 50(1):382–384.

Brooks MJ, Laurence CA, Hansen EJ, Gray-Owen SD. 2007. Characterization of the *Moraxella catarrhalis* Opa-like protein, OlpA, reveals a phylogenetically conserved family of outer membrane proteins. J Bacteriol. 189:76–82.

Chan JZ, Halachev MR, Loman NJ, Constantinidou C, Pallen MJ. 2012. Defining bacterial species in the genomic era: insights from the genus *Acinetobacter*. BMC Microbiol. 12:302.

D'Argenio DA, Miller SI. 2004. Cyclic di-GMP as a bacterial second messenger. Microbiology 150:2497–2502.

Darling AC, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res. 14:1394–1403.

Davie JJ, et al. 2011. Comparative analysis and supragenome modeling of twelve *Moraxella catarrhalis* clinical isolates. BMC Genomics 12:70.

de Vries SP, Bootsma HJ, Hays JP, Hermans PW. 2009. Molecular aspects of *Moraxella catarrhalis* pathogenesis. Microbiol Mol Biol Rev. 73:389–406.

de Vries SP, Eleveld MJ, Hermans PW, Bootsma HJ. 2013. Characterization of the molecular interplay between *Moraxella catarrhalis* and human respiratory tract epithelial cells. PLoS One 8(8):e72193.

de Vries SP, et al. 2010. Genome analysis of *Moraxella catarrhalis* strain BBH18, a human respiratory tract pathogen. J Bacteriol. 192:3574–3583.

de Vries SP, et al. 2014. Deciphering the genetic basis of *Moraxella catarrhalis* complement resistance: a critical role for the disulphide bond formation system. Mol Microbiol. 91:522–537.

Dingman JR, et al. 1998. Correlation between presence of viable bacteria and presence of endotoxin in middle-ear effusions. J Clin Microbiol. 36:3417–3419.

Donati C, et al. 2010. Structure and dynamics of the pan-genome of *Streptococcus pneumoniae* and closely related species. Genome Biol. 11:R107.

Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol. 7:214.

Edwards KJ, et al. 2005. Multiplex PCR assay that identifies the major lipooligosaccharide serotype expressed by *Moraxella catarrhalis* clinical isolates. J Clin Microbiol. 43:6139–6143.

Faden H. 2001. The microbiologic and immunologic basis for recurrent *otitis media* in children. Eur J Pediatr. 160:407–413.

Goris J, et al. 2007. DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. Int J Syst Evol Microbiol. 57:81–91.

Green RE, et al. 2010. A draft sequence of the Neanderthal genome. Science 328:710–722.

Hall BG, Ehrlich GD, Hu FZ. 2010. Pan-genome analysis provides much higher strain typing resolution than multi-locus sequence typing. Microbiology 156:1060–1068.

Hall-Stoodley L, et al. 2006. Direct detection of bacterial biofilms on the middle-ear mucosa of children with chronic otitis media. JAMA 296:202–211.

Hill DJ, Virji M. 2003. A novel cell-binding mechanism of *Moraxella catarrhalis* ubiquitous surface protein UspA: specific targeting of the N-domain of carcinoembryonic antigen-related cell adhesion molecules by UspA1. Mol Microbiol. 48:117–129.

Hill DJ, Whittles C, Virji M. 2012. A novel group of *Moraxella catarrhalis* UspA proteins mediates cellular adhesion via CEACAMs and vitronectin. PLoS One 7:e45452.

Hiller NL, et al. 2007. Comparative genomic analyses of seventeen *Streptococcus pneumoniae* strains: insights into the pneumococcal supragenome. J Bacteriol. 189:8186–8195.

Hoban DJ, Doern GV, Fluit AC, Roussel-Delvallez M, Jones RN. 2001. Worldwide prevalence of antimicrobial resistance in *Streptococcus pneumoniae*, *Haemophilus influenzae*, and *Moraxella catarrhalis* in the SENTRY Antimicrobial Surveillance Program, 1997-1999. Clin Infect Dis. 32(Suppl. 2):S81–S93.

Hogg JS, et al. 2007. Characterization and modeling of the *Haemophilus influenzae* core and supra-genome based on the complete genomic sequences of Rd and 12 clinical nontypeable 16 strains. Genome Biol. 8:R103.

Huson DH, Bryant D. 2006. Application of phylogenetic networks in evolutionary studies. Mol Biol Evol. 23:254–267.

Johnson DM, Sader HS, Fritsche TR, Biedenbach DJ, Jones RN. 2003. Susceptibility trends of *Haemophilus influenzae* and *Moraxella catarrhalis* against orally administered antimicrobial agents: five-year report from the SENTRY Antimicrobial Surveillance Program. Diagn Microbiol Infect Dis. 47:373–376.

Konstantinidis KT, Ramette A, Tiefje JM. 2006. The bacterial species definition in the genomic era. Philos Trans R Soc Lond B Biol Sci. 361:1929–1940.

Konstantinidis KT, Tiedje JM. 2005. Genomic insights that advance the species definition for prokaryotes. Proc Natl Acad Sci U S A. 102:2567–2572.

Ladner JT, Whitehouse CA, Koroleva GI, Palacios GF. 2013. Genome sequence of *Moraxella macacae* 0408225, a novel bacterial species isolated from a cynomolgus macaque with epistaxis. Genome Announc. 1(1): e00188–12.

Marshall DJ, et al. 1997. Determination of hepatitis C virus genotypes in the United States by cleavase fragment length polymorphism analysis. J Clin Microbiol. 35:3156–3162.

Melendez PR, Johnson RH. 1991. Bacteremia and septic arthritis caused by *Moraxella catarrhalis*. Rev Infect Dis. 13:428–429.

Murphy TF, Brauer AL, Grant BJ, Sethi S. 2005. *Moraxella catarrhalis* in chronic obstructive pulmonary disease: burden of disease and immune response. Am J Respir Crit Care Med. 172:195–199.

Murphy TF, Parameswaran GI. 2009. *Moraxella catarrhalis*, a human respiratory tract pathogen. Clin Infect Dis. 49:124–131.

Myers EW, et al. 2000. A whole-genome assembly of *Drosophila*. Science 287:2196–2204.

Nistico L, et al. 2009. Fluorescence "in situ" hybridization for the detection of biofilm in the middle ear and upper respiratory tract mucosa. Methods Mol Biol. 493:191–213.

Nistico L, et al. 2014. Using the core and supra genomes to determine diversity and natural proclivities among bacterial strains. In: Skovhus TL, Caffrey S, Hubert C, editors. Molecular microbial methods. Norfolk (UK): Caister Academic Press.

Noonan JP, et al. 2006. Sequencing and analysis of the Neaderthal genomic DNA. Science 314:1113–1118.

Peak IR, et al. 2007. Towards understanding the functional role of the glycosyltransferases involved in the biosynthesis of *Moraxella catarrhalis* lipooligosaccharide. FEBS J. 274:2024–2037.

Pearson WR, Lipman DJ. 1988. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A. 85:2444–2448.

Perez AC, et al. 2014. Residence of *Streptococcus pneumoniae* and *Moraxella catarrhalis* within polymicrobial biofilm promotes antibiotic resistance and bacterial persistence *in vivo*. Pathog Dis. 70(3):280–288.

Pickford M, Andrews P. 1981. The Tinderet Miocence sequence in Kenya. J Hum Evol. 10:11–33.

Pingault NM, Lehmann D, Bowman J, Riley TV. 2007. A comparison of molecular typing methods for *Moraxella catarrhalis*. J Appl Microbiol. 103:2489–2495.

Plamondon P, Luke NR, Campagnari AA. 2007. Identification of a novel two-partner secretion locus in *Moraxella catarrhalis*. Infect Immun. 75:2929–2936.

Post JC, et al. 1995. Molecular analysis of bacterial pathogens in otitis media with effusion. JAMA 273:1598–1604.

Richter M, Rosello-Mora R. 2009. Shifting the genomic gold standard for the prokaryotic species definition. Proc Natl Acad Sci U S A. 106:19126–19131.

Riley MA, Lizotte-Waniewski M. 2009. Population genomics and bacterial species concept. Methods Mol Biol. 532:368–377.

Shen K, et al. 2005. Identification, distribution, and expression of novel (nonRd) genes in ten clinical isolates of nontypeable *Haemophilus influenzae*. Infect Immun. 73:3479–3491.

Shen K, et al. 2006. Characterization, distribution and expression of novel genes among eight clinical isolates of *Streptococcus pneumoniae*. Infect Immun. 74(1):321–330.

Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 30:1312–1313.

Stol K, et al. 2012. Inflammation in the middle ear of children with recurrent or chronic *otitis media* is associated with bacterial load. Pediatr Infect Dis J. 31:1128–1134.

Su YC, Singh B, Riesbeck K. 2012. *Moraxella catarrhalis*: from interactions with the host immune system to vaccine development. Future Microbiol. 7:1073–1100.

Vaneechoutte M, Verschraegen G, Claeys G, Flamen P. 1988. Rapid identification of *Branhamella catarrhalis* with 4-methylumbelliferyl butyrate. J Clin Microbiol. 26:1227–1228.

Vaneechoutte M, Verschraegen G, Claeys G, Van Den Abeele AM. 1990. Serological typing of *Branhamella catarrhalis* strains on the basis of lipopolysaccharide antigens. J Clin Microbiol. 28:182–187.

Verduin CM, Hol C, Fleer A, van Dijk H, van Belkum A. 2002. *Moraxella catarrhalis*: from emerging to established pathogen. Clin Microbiol Rev. 15:125–144.

Verhaegh SJ, et al. 2011. Colonization of healthy children by *Moraxella catarrhalis* is characterized by genotype heterogeneity, virulence gene diversity and co-colonization with *Haemophilus influenzae*. Microbiology 157:169–178.

Verhaegh SJC, et al. 2008. Age-related genotypic and phenotypic differences in *Moraxella catarrhalis* isolates from children and adults presenting with respiratory disease in 2001-2002. Microbiology 14:1178–1184.

Wang W, et al. 2007. Metabolic analysis of *Moraxella catarrhalis* and the effect of selected *in vitro* growth conditions on global gene expression. Infect Immun. 75:4959–4971.

Wirth T, et al. 2007. The rise and spread of a new pathogen: seroresistant *Moraxella catarrhalis*. Genome Res. 17:1647–1656.

Yahara K, et al. 2014. Efficient inference of recombination hot regions in bacterial genomes. Mol Biol Evol. 31:1593–1605.

Yeoman CJ, et al. 2010. Comparative genomics of *Gardnerella vaginalis* strains reveals substantial differences in metabolic and virulence potential. PLoS One 5:e12411.

Zomer A, et al. 2012. Genome sequence of *Moraxella catarrhalis* RH4, an isolate of seroresistant lineage. J Bacteriol. 194:6969.

Associate editor: Tal Dagan

Genome Biology and Evolution

SMBE